AI Resilience: Planning for Absence, Not Degradation

Yasin Jabbar · yasinjabbar.com · September 2026

Primary sources (all public, linked at the end): IBM Institute for Business Value, *The Calculus of AI Sovereignty* (17 June 2026) · Zapier enterprise AI vendor lock-in survey (2 April 2026) · OpenAI, Anthropic and xAI public status pages for 3 September 2026, as reported by Wired and Engadget · Microsoft Online Services SLA for Azure OpenAI Service · Cloud Security Alliance, *AI Provider Concentration Risk: Enterprise Resilience* (June 2026). Nothing in this chapter draws on non-public information from any prior employer. General industry practice only.

The Decision

On the morning of Thursday, 3 September 2026, three of the largest model providers in the world went dark inside the same ninety minutes. Anthropic's Claude started throwing elevated errors at 9:26 a.m. Eastern. xAI's Grok followed at about 9:30. OpenAI's ChatGPT and Codex went at 10:43. The causes were unrelated — a routing error at one, an infrastructure fault at another, a data-center failure at the third — and by late afternoon all three were back.

Nobody planned it. Nobody coordinated it. And for a few hours, a large share of the coding agents, support bots, document pipelines and internal copilots in American business simply stopped.

Here is the thing to notice: it wasn't a slowdown. When a model API goes dark, an AI-dependent workflow does not degrade gracefully the way a slow database does. It halts. The claim never gets triaged. The ticket never gets a first response. The overnight batch that an agent was supposed to run at 2 a.m. does not run. Then, hours later, everything comes back and there is no incident report because — as far as your systems are concerned — nothing you own ever failed.

That is the decision in front of every mid-market CIO right now. Not *whether* to depend on a model provider. You already do. The decision is **how to be absent-proof**: what happens to the business on the day the model you built on is not there, and who signed off on that answer before it happened.

Decision 3 argued that accountability has three layers — visibility, containment, decision rights — and put continuity inside the containment layer as one line item. This chapter is that line item, opened up. Because the data says it is the one most organizations have quietly skipped.

The Pain — In Their Own Numbers

IBM's Institute for Business Value, working with Oxford Economics, surveyed 1,000 senior executives responsible for AI, data or technology across 16 countries and 17 industries between February and April 2026. The headline findings from *The Calculus of AI Sovereignty*:

- **91%** say they do not fully understand their organization's dependencies across AI vendors, models and infrastructure.
- Surveyed leaders report an average of **six AI-related disruptions in the past two years**, largely driven by vendor services.
- **81%** say a seven-day vendor outage would cause severe or critical disruption — effectively halting operations.
- **71%** say switching their primary AI vendor or model would be difficult.

- Only **7%** operate at the most advanced level of AI control — and that minority protects **55% more operating profit** from AI-driven disruptions than everyone else.
- **72%** would accept a 20% cost increase to keep a vendor if it improved strategic flexibility.

Read the first three together. Three disruptions a year on average, most of them caused by somebody else's service; the majority admit they cannot map where the dependencies are; and four in five say a week without the vendor would stop the business. That is not a risk register. That is a confession.

Zapier surveyed 542 U.S. executives with paid AI vendor contracts in early 2026 and found the same shape from a different angle:

- **74%** say losing their primary AI vendor would disrupt day-to-day operations or leave them unable to function. Only **6%** could walk away with no disruption.
- **89%** believe they could switch vendors within four weeks. Of the two-thirds who have actually tried a migration, **58%** say it failed or took far more effort than expected.
- **44%** run multiple vendors; **42%** maintain a contingency plan. Which means more than half do neither.

The gap between 89% who think they could switch in a month and 58% who tried and could not is the whole story of this chapter. Confidence about resilience is running about two years ahead of actual resilience.

The False Choice — "Trust the SLA" vs. "Multi-Vendor Everything"

The two default answers are both wrong, for opposite reasons.

"We're covered — the contract has a 99.9% SLA." The published Azure OpenAI Service SLA is a 99.9% *monthly uptime* commitment. Do the arithmetic: 99.9% permits roughly 44 minutes of downtime a month, about 8.8 hours a year. The remedy for a breach is a service credit — 10% of the affected resource's monthly fee if uptime falls below 99.9%, 25% below 99%, 100% below 95%. A four-hour outage that stops your claims desk earns you a coupon worth a tenth of that month's token bill.

And the exclusions are where the real failures live. Uptime means the endpoint answered. It does not mean it answered quickly, it does not mean the content filter was working, it does not mean the model version you tuned against is still available, and a burst of 429 "too many requests" responses during a capacity crunch counts as *up*. You also carry the burden of proof: no telemetry from your side, no claim.

OpenAI's own contractual uptime commitments apply to specific enterprise and scale tiers; below those tiers there is no contractual guarantee at all. Anthropic's enterprise commitments are negotiated per customer and not published. None of this is a vendor failing. It is what a young, capacity-constrained industry can honestly promise. The mistake is treating the promise as a continuity plan.

"We'll go multi-vendor and route around anything." Also wrong, in a subtler way. On 3 September, an organization with fallback from OpenAI to Anthropic to xAI would have watched all three fail inside ninety minutes for three unrelated reasons. Multi-vendor buys you independence from one company's bad day. It does not buy you independence from the category. And IBM's data shows that most "multi-vendor" estates (73% describe themselves that way) got there by accident — independent business-unit purchases (69%) and geographic necessity (69%) — not by design. Two vendors nobody planned is not redundancy. It is two dependencies.

The real decision is neither. It is deciding, workload by workload, **how long you can be without the model, and building exactly enough to survive that long.**

The Real Decision — Three Tiers of Absence

Every AI-dependent workload in the building belongs in one of three tiers. The tier is set by a single question: *if the model is gone for four hours, what happens to revenue, customers or compliance?*

Tier	If the model is absent for 4 hours…	Required posture
A — Stops the business	Revenue, customer commitments or a regulatory clock are directly affected.	Tested manual fallback **and** a second technical path. Named owner. Drilled quarterly.
B — Degrades the business	Work queues up; cost and customer-experience damage accrue; nothing is legally or contractually broken.	Documented manual fallback, tested twice a year. Queue-and-replay design so nothing is lost.
C — Inconveniences people	Internal productivity dips. Nobody outside the company notices.	Inventory entry and an owner. Nothing more. Do not spend money here.

Most organizations, when they run this exercise for the first time, discover two things. First, that they have more Tier A workloads than they thought, because an agent someone stood up in an afternoon has quietly become load-bearing (Decision 3's orphaned-system pattern). Second, that they have been spending their resilience budget on Tier C — dual-vendor contracts for the internal chatbot — while Tier A runs on a single API key and hope.

Tier A is where the money goes. For those workloads, "second technical path" means one of three things, in ascending order of cost: a second provider behind an abstraction layer you actually test; a smaller model you host yourself that is good enough to keep the lights on; or a non-AI procedure — a rules engine, a form, a phone number — that produces an acceptable outcome slowly. The third option is unfashionable and, for most mid-market Tier A workloads, the right answer. The claims desk does not need a second frontier model. It needs the triage rules it used in 2023 kept current and a way to switch to them in ten minutes.

Queue-and-replay is the cheapest resilience there is. For Tier B, the design principle is that an absent model should cost you time, never data. Every request the model would have handled goes into a durable queue; when the provider returns, the queue drains. A workflow that drops requests on the floor during an outage has a data-loss problem disguised as an availability problem.

Absence is broader than outages. IBM's respondents cited price increases, usage restrictions, model deprecations and performance degradation alongside downtime. A model version you tuned against being retired on 90 days' notice is an absence event. A 3x price increase at renewal is an absence event, slower. The tiering holds for all of them: Tier A workloads need a written answer to "what if this model is not available at this price in twelve months," and the answer is not "we'll negotiate."

The 90-Day Cure

This is a habit with three artifacts, run in three sprints.

Days 1–30 — Inventory and tier. One spreadsheet. Every workload with a model in the path: which provider, which model version, who owns the relationship, which tier, what breaks at four hours, what breaks at seven days. You will find things. The IBM 91% is not an insult; it is a description of everyone's first inventory. Output: a tiered dependency map with a name on every row, signed by the CIO.

Days 31–60 — Fallbacks and the first drill. For every Tier A row, write the manual procedure and identify the second path. Then kill the model in staging — revoke the key, black-hole the endpoint — and watch what actually stops. The drill is the point. A fallback that has never been exercised is a document, not a fallback. Output: a drill report, ugly and honest, with the fixes it produced.

Days 61–90 — Contracts and telemetry. Take the tiered map to procurement. For Tier A providers, negotiate what the public SLA does not give you: a latency floor, capacity guarantees, a model-availability commitment (the version you depend on stays available for an agreed period), a named escalation contact, and automatic credits. Buyers with meaningful committed spend get these; buyers who do not ask get the public floor. Stand up the telemetry that proves a breach — external synthetic probes, request logs with timestamps, 429 and 5xx counts by deployment — because without it, even the credits you are owed will not be paid. Output: an SLA gap list per Tier A vendor, and a monitoring dashboard that would have caught 3 September before the users did.

After day 90 it becomes a cadence: the inventory is reviewed monthly, Tier A is drilled quarterly, contracts are reopened twelve months before renewal, not at signature.

Common Wrong Turns

- **"We'll build an abstraction layer so we can swap models any time."** Useful, and not sufficient. The abstraction layer handles the API call; it does not handle the prompts, evaluations, fine-tuning and workflow logic tuned to one model's behavior. That is why 58% of attempted migrations disappointed. Build the layer, then test the swap end-to-end on a real workload before you count it.
- **"Multi-region failover solves it."** It solves *one provider's* regional failure, at a cost most deals never budgeted — active-active roughly doubles provisioned spend; warm standby adds 30–50%. It does nothing for a provider-wide routing error, and nothing for a deprecation or a price increase.
- **"Our AI vendor is a hyperscaler, so it's four-nines."** The compute substrate may be. The model service on top of it is contracted at 99.9% in public, and the exclusions are what will bite you. Read the SLA for the thing you actually call, not the platform beneath it.
- **"We'll worry about it when we have more AI in production."** The IBM average is three vendor-driven disruptions a year *now*. The workloads you have already are the ones that will fail first, and the inventory takes a month.
- **"Resilience is the security team's job."** It is a continuity job, and it belongs to whoever owns the workload's revenue. Security can tell you a key was revoked. Only the business owner can tell you whether the fallback outcome was acceptable.

What This Costs to Run

The play — a tiered inventory, a quarterly Tier A drill, and a contract-plus-telemetry pass on the vendors that matter — is roughly the same half-engineer habit as Decision 3, plus a modest and *deliberately targeted* spend on second paths for Tier A only. For most mid-market organizations that is a handful of workloads, and for several of them the second path is a procedure, not a product.

On the other side of the ledger: IBM's respondents say a seven-day outage halts the business for 81% of them; the 7% with real control protect 55% more operating profit from exactly these events; and 72% of their peers say they would pay a 20% premium for the flexibility that this play produces for a fraction of that. You do not need a consultant's multiplier. Ask your CFO what one Tier A workload stopping for a business day costs, and compare it to the price of keeping the 2023 procedure current.

Recommendation by Profile

- **Regulated mid-market** (healthcare, financial services, credit unions and community banks, gov-adjacent): Full 90-day cure, with the drill report going to the risk committee. Your examiners already ask this question about core banking and claims platforms; they will ask it about the model behind your member-service bot next. Be the institution with the answer written down.

- **Growth-stage SaaS:** Tier the inventory in 30 days; second technical path for every Tier A workload that touches a customer commitment in 60; queue-and-replay as a standard pattern for everything new. Your customers' SLAs to *their* customers are now downstream of your model provider's. Know that before they do.
- **Cost-sensitive operations:** Inventory and tiering only, then manual fallback procedures for Tier A. No second-vendor spend in year one. The cheapest resilience is a documented way to do the work by hand for a day, and most operations shops already know how.

The Operator's Bottom Line

> **The provider will be back in four hours. The question is whether your business will be — and whether anyone can say, in writing, what it did in the meantime.**

Resilience for AI is not a product you buy and not a clause you negotiate. It is a tiered list of what you depend on, a tested way to be without it, and a contract that reflects what you learned. Three artifacts. Ninety days. Plan for absence, not degradation — then drill it before the calendar does it for you.

Sources

1. IBM Institute for Business Value, "IBM Study: Limited Control and Rising Dependencies Leave Enterprises Exposed in the Age of AI," press release, 17 June 2026 (*The Calculus of AI Sovereignty*, 1,000 senior executives, 16 countries, 17 industries, with Oxford Economics). https://newsroom.ibm.com/2026-06-17-ibm-study-limited-control-and-rising-dependencies-leave-enterprises-exposed-in-the-age-of-ai

2. Zapier, "Zapier Survey Finds Nearly 3 in 4 Enterprises Would Face Disruption If They Lost Their Primary AI Vendor," Business Wire, 2 April 2026 (542 U.S. executives with paid AI vendor contracts, surveyed 30 Jan–6 Feb 2026). https://www.businesswire.com/news/home/20260402086941/en/

3. OpenAI status incident, 3 September 2026 ("Elevated errors across ChatGPT and Codex"), https://status.openai.com/incidents/01M1KWEDH417T2CF44YYHZDFCR · Anthropic status history, https://status.claude.com/ · xAI status history, https://status.x.ai/grok-com

4. Wired, "Nobody Is Saying Why OpenAI and Anthropic Had Outages Today," 3 September 2026. https://www.wired.com/story/nobody-is-saying-why-openai-and-anthropic-had-outages-today/

5. Engadget, "xAI apologizes for outage that affected Grok and other compute partners," September 2026. https://www.engadget.com/2250789/spacexai-apologizes-for-outage-that-affected-grok-and-other-compute-partners/

6. Microsoft, Online Services Service Level Agreements (Azure OpenAI Service: 99.9% monthly uptime; tiered 10% / 25% / 100% credits). https://www.microsoft.com/licensing/docs/view/Service-Level-Agreements

7. Cloud Security Alliance, *AI Provider Concentration Risk: Enterprise Resilience*, June 2026. https://labs.cloudsecurityalliance.org/research/ai-provider-concentration-risk-enterprise-resilience-v1-csa/

Next in the series — Decision 5: "Governance for Shadow AI: How to Stop Punishing the Smartest People in Your Building." Why mandatory AI training is the fastest way to grow the shadow estate, and what to build instead.