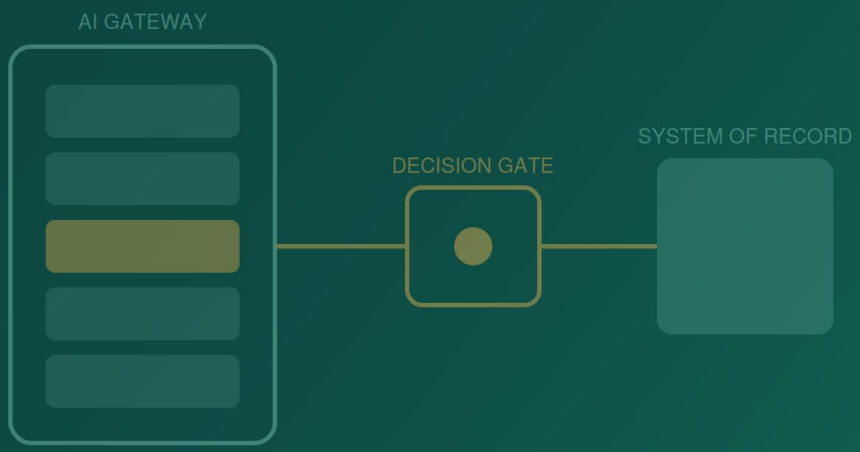
SOVEREIGN DIGITAL RESILIENCE · PAPER 2

Governing Agentic AI

In Systems of Public Record, Governance Is the Accelerant

A practitioner's operating model for agentic AI where the authoritative write is the crown jewel.



Sanjeet Kumar

Enterprise Architecture & Technology Governance Leader

sanjeetkumar.com/papers/governing-agentic-ai · 2026 · Views are the author's own.

SOVEREIGN DIGITAL RESILIENCE · WORKING PAPER

Governing Agentic AI

In Systems of Public Record, Governance Is the Accelerant

Sanjeet Kumar — Enterprise Architecture & Technology Governance Leader

Critical Digital Infrastructure · Sovereign Digital Resilience · AI Governance — 2026 · Views are the author's own.

ORCID 0009-0003-1677-684X · DOI: 10.5281/zenodo.21971009

A practitioner's operating model for agentic AI where the authoritative write is the crown jewel.

All deposits and versions: sanjeetkumar.com/papers/governing-agentic-ai

The Demo and the Ledger

Every agentic-AI initiative has two moments. The first is euphoric: someone announces, in effect, '*I have found the silver/AI bullet*'. The second arrives after the demo has gone well, and it is the one I have come to watch for. An agent has just read a backlog item, traced the relevant records, drafted the correction, written the test, and prepared the change — in minutes, unassisted. Someone in the room says what everyone is thinking: *"So… can we let it run?"*

In most software, that is a product question. In a system of public record — a land registry, a core-banking ledger, a civil-status database — it is a constitutional one. These systems exist to be the authoritative answer to a question citizens keep asking: *who owns this, who owes what, what is true of record?* Their value is not the software; it is the warrant behind every write. An agent that accelerates work around such a system is a gift. An agent that reaches the point of record without a governing architecture is a liability being amortized — every impressive demo adding to a balance that comes due on the day the answer to "what happens when it's wrong?" turns out to be "we're not sure, and nobody specifically."

The question I am asked most often about agentic AI is "what can it do?" In systems of public record, that is the wrong first question. The right first question is: **what happens when it's wrong, and who is accountable?** Answer that architecturally — before the first agent is provisioned — and aggressive adoption becomes safe. Skip it, and governance arrives later anyway, as an incident review.

My team had the opportunity to initiate and shape an agentic-AI modernization MVP inside exactly such an environment — an initiative that delivered acknowledged multi-million-dollar savings and cut MVP cycles from roughly six months to six weeks. This paper is the operating model behind that result, expanded from the version I first published, because the most interesting part of the story is still the least intuitive: the acceleration came because of the governance, not despite it. It is also a companion piece: in The AI Gateway Advantage I argued that an organization-owned gateway is how you govern the inference path — what models are called, at what cost, with what telemetry. This paper extends the same control plane to the action path: what agents are permitted to do, and how you prove it.

The Threat Model Nobody Writes Down

Traditional application threat models assume the untrusted party is outside the perimeter. Agentic systems break that assumption politely: the most consequential untrusted actor is now a component 'we deployed' on purpose. Four failure modes matter most in systems of record, and none of them is exotic:

Hallucinated authorization. An agent convinces itself — or a downstream system — that it holds permissions it does not. This is not malice; it is a fluent system completing a pattern. If the only thing standing between an agent and a write is the agent's own understanding of its role, you have no boundary at all.

The confused deputy. An agent with legitimate privileges is manipulated into exercising them for an illegitimate purpose — the classic security failure, now reachable through natural language. Instructions embedded in retrieved documents, user-supplied content, or tool outputs become attack surface, a risk class the OWASP LLM and agentic-security work has now catalogued thoroughly.

Compounding autonomy. Chains of individually reasonable delegations — agent calls tool, tool calls service, service triggers workflow — that no one reviewed as a whole. Each hop was approved; the composition never was. Blast radius is a property of the chain, not the link.

Provenance decay. Six months later, a record was changed and the organization cannot reconstruct why: which model, which prompt version, which retrieved context, which human approved it. In ordinary software this is an annoyance. In a registry, where the record must outlive every system that touched it, it is an integrity failure — the same century-scale test I applied to data sovereignty in Century-Scale Systems applies to decisions about the record, not just the record itself.

Notice what these have in common: none is prevented by a better prompt. All are prevented by architecture. That observation drives everything that follows.

Four Principles

1. Draw the authority boundary before the capability boundary.

The first architectural artefact of our MVP was not a model selection — it was a decision map: which actions the agent may execute autonomously, which require human confirmation, and which are categorically reserved for humans. In a system of public record, the authoritative write is the crown jewel; our guardrail was simple and absolute — *agents accelerate everything up to the point of record, and an accountable human owns the point of record.* Autonomy is granted per **action class**, not per system: the same agent may run free on test generation and be firmly gated on record mutation. Capability is what the vendor demos. Authority is what our architecture grants. Confusing the two is how organizations end up with an org chart that quietly contains software.

2. Treat the agent as an untrusted, talented junior.

Bounded workspaces. Least-privilege, short-lived credentials. Outputs that are proposals until validated. We would extend exactly this structure to a brilliant new hire on their first week — trust in their talent, verification of their work, and no production keys. Hallucinated authorization is prevented structurally, at the permission layer, not behaviourally, in the prompt. **Prompts are advisory; credentials are architecture.** The corollary for the confused deputy: an agent's *inputs* are untrusted too, so anything retrieved or received is treated as data, never as instruction, and tool scopes are narrow enough that a manipulated agent is a contained agent.

3. Extend the architecture repository to cover agents.

Traditional EA artefacts describe systems, interfaces, and data flows. Agentic systems require the repository to also carry: agent permission matrices, human-in-the-loop checkpoints, model and prompt lineage, and decision provenance — who or what recommended each action, on what context, approved by whom. This is TOGAF-style discipline extended to a new class of actor, and it is what turns “we use AI” from a claim into an auditable statement. It is also, increasingly, what regulators mean when they ask for AI governance: NIST's AI Risk Management Framework, ISO/IEC 42001, and Canada's Directive on Automated Decision-Making converge on the same demand — show your decision rights, show your records.

4. Sovereignty applies to the whole inference chain.

Prompts, retrieval context, embeddings, and outputs inherit the sensitivity of the records they touch. If a citizen's record flows through an inference endpoint, that endpoint sits inside the sovereignty boundary — subject to the same residency, jurisdictional-control, and audit expectations as the database it came from. Regulatory regimes — the UAE's PDPL, Canadian privacy law, sectoral outsourcing rules — are converging on this position faster than most architectures are. Sovereignty is not a paragraph in a policy document; as I argued in the gateway paper, it is an **enforced routing policy** — and agents, which assemble context automatically and promiscuously, make enforcement more necessary, not less.

The Control Plane: Where the Gateway Meets the Agent

I live and breathe architecture principles. Principles need an enforcement point, and here the two papers join. The organization-owned AI gateway I described in The AI Gateway Advantage paper was built to govern the inference path: a unified API in front of every model, routing and tiering, policy and guardrails, telemetry that measures value rather than spend, budget controls, and an audit log. The critical realization of the agentic era is that the same gateway is the natural control plane for the action path. Nothing an agent does — no model call, no tool invocation that matters — should bypass it.

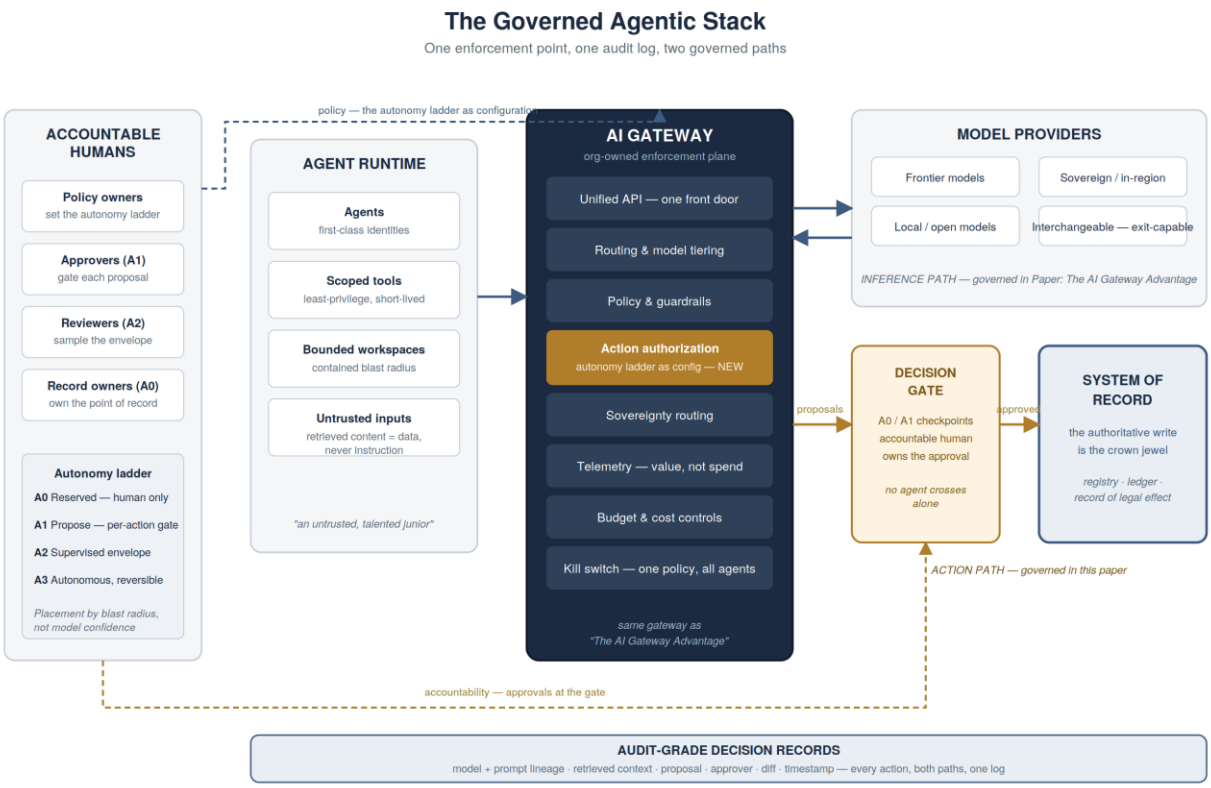


Figure 1 — The Governed Agentic Stack. *The gateway built for inference governance becomes the control plane for action governance. The upper flow is the subject of The AI Gateway Advantage; the lower flow is the subject of this paper. One enforcement point, one audit log, two paths.*

Read the diagram bottom-up from the asset being protected. The **system of record** sits behind a decision gate no agent crosses alone. The **gateway** enforces, in one place: identity for agents (every agent a first-class principal, never a shared service account), action policy (the autonomy ladder below, expressed as configuration rather than convention), sovereignty routing (which data classes may reach which endpoints, in which jurisdictions), value telemetry, and the decision record. Above it, the **agent runtime** — agents, tools, bounded workspaces — connects to models on one path and proposes actions on the other. Beside everything, **accountable humans** hold the approval tiers.

Three properties make this arrangement more than tidy:

One choke point, honestly held. Policies enforced in one owned component are policies that are actually enforced. Guardrails distributed across a dozen agent frameworks are guardrails you hope for. This is the de-perimeterization lesson applied inward — and its commercial twin from Exit-Capable by Design applies too: because the gateway abstracts providers, the governance survives any vendor exit. You can change models without renegotiating your controls.

The kill switch is real. When every agent action transits the gateway, "pause all autonomous actions" is one policy change, effective in seconds, with a record of what was in flight. Ask any incident commander what that is worth.

Telemetry becomes governance evidence. The same instrumentation that told us which teams generated value per token now tells auditors which actions were proposed, gated, approved, and executed — for free, because it is the same pipe.

Decision Rights: The Autonomy Ladder

The decision map in Principle 1 needs a vocabulary. We used four action classes, and I have yet to find a system of record that needs more:

Class	Name	Agent may…	Employee role	Typical actions	Evidence captured
A0	Advisory	Draft and recommend only — never execute	Owns the action entirely	Authoritative writes, legal attestations, anything altering the record's warrant	Full provenance of the recommendation
A1	Propose	Prepare the complete action	Approves each instance before execution	Record corrections, customer-visible outputs, config changes	Proposal, approver identity, diff, timestamp
A2	Supervised	Execute within a pre-approved envelope	Samples and reviews; owns the envelope	Classification, metadata enrichment, batch analysis	Envelope definition, execution log, sample results
A3	Autonomous	Execute freely; actions reversible by design	Sets policy; audits on exception	Test generation, draft documents, triage, research	Action log, rollback path

Two rules give the ladder its force. **Placement is by blast radius and reversibility, not by confidence in the model** — a better model next quarter does not promote an action class; a demonstrated rollback path might. And **movement is one-way gated**: promoting an action class from A1 to A2 is an architecture decision with a named owner and a written rationale in the repository — never a config drift. The ladder is also where the accelerant lives, which brings me to the result.

What Actually Accelerated

Here is the sequence sceptics expect: governance framework, six months of committee, then a cautious pilot. Here is what happened instead: because the authority boundary was drawn *first*, everything below the point of record became an explicit safe zone — and inside a safe zone, engineers stop asking permission. The negotiation that normally strangles AI initiatives (“is this allowed?”, asked case-by-case, answered by whoever is nervous that week) had already been answered once, architecturally, for everyone. A0/A1 gates protected the record; A2/A3 freedom compounded weekly. MVP cycles fell from roughly six months to six weeks not despite the gates but because the gates made the open ground legible.

There is a second-order effect I did not predict. Explicit governance made the initiative politically fast: risk, audit, and business owners approved expansion quickly because every approval request arrived with its decision class, its blast radius, and its evidence trail attached. Trust, it turns out, compounds faster than

capability. Teams with the fewest rules move fast for a quarter; teams whose rules are architectural move fast for years.

The Convergence

If the practitioner's case does not persuade, the regulatory one will arrive on schedule. NIST's AI RMF and its generative-AI profile, ISO/IEC 42001's management-system requirements, the EU AI Act's obligations for high-risk systems, Canada's Directive on Automated Decision-Making with its tiered human-in-the-loop requirements, and the UAE's PDPL alongside its National AI Strategy 2031 differ in vocabulary and reach — but they are converging on precisely the artefacts this operating model produces as a by-product: documented decision rights, human accountability proportionate to impact, provenance of automated decisions, and jurisdictional control of the data chain. Organizations that governed early will comply by printing reports. Organizations that didn't will retrofit under supervision.

Calibrated to What the System Can Afford to Get Wrong

The organizations that win the agentic era will not be those with the fewest rules, but those whose rules are architectural: enforced by design at an owned control plane, auditable by default, and calibrated to what the system can afford to get wrong. In systems of public record the calibration is strict, because the asset is not data — it is the public's warrant that the record is true. Agents can accelerate everything that surrounds that warrant, and should. The point of record stays human, and the architecture — not the prompt — is what makes that promise real.

In systems societies can't afford to lose, that calibration is the job. Some of us have been doing it for years; agentic AI has simply made the job visible.

References

1. National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, January 2023.
2. National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1, July 2024.
3. International Organization for Standardization, ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system, 2023.
4. European Union, Regulation (EU) 2024/1689 (Artificial Intelligence Act), Official Journal of the European Union, 2024.
5. Treasury Board of Canada Secretariat, Directive on Automated Decision-Making, Government of Canada.
6. OWASP Foundation, OWASP Top 10 for Large Language Model Applications, 2025 edition.
7. OWASP Foundation, Agentic Security Initiative, Agentic AI — Threats and Mitigations, 2025.
8. Anthropic, Building Effective Agents, December 2024.
9. United Arab Emirates, Federal Decree-Law No. 45 of 2021 on the Protection of Personal Data (PDPL).

10. United Arab Emirates, UAE National Strategy for Artificial Intelligence 2031.

11. The Open Group, The TOGAF® Standard, 10th Edition, 2022.

12. Kumar, S., The AI Gateway Advantage: Value over Spend. Smart Work over Hard Work., sanjeetkumar.com/papers/ai-gateway-advantage, 2026. doi:10.5281/zenodo.21970942

13. Kumar, S., Century-Scale Systems: What Land Registries Teach Us About Data Sovereignty, sanjeetkumar.com/papers/century-scale-systems, 2026. doi:10.5281/zenodo.21970840

14. Kumar, S., Exit-Capable by Design: Vendor Concentration Is a Sovereignty Risk, sanjeetkumar.com/papers/exit-capable-by-design, 2026. doi:10.5281/zenodo.21970869