WHITEPAPER

Controlling agent actions: a reference architecture

A vendor-neutral reference architecture for deciding, enforcing and recording what AI agents may do, with a maturity model and a mapping to the EU AI Act, ISO/IEC 42001 and NIST AI RMF.

By Your Name

Published 30 October 2026 · 3 min read

Runtime policy enforcement | Agent identity and access | Observability and audit | Regulation and standards

KEY FINDINGS

1. A single enforcement point in front of every tool call is the control that most other controls depend on; without it, logging and approvals have nothing reliable to attach to.
2. Agents should hold their own identities with task-scoped, short-lived credentials. Sharing a human user's session makes actions unattributable and permissions impossible to narrow.
3. Tracking whether untrusted content influenced a request, and tightening policy when it did, reduces exposure to prompt injection more than input filtering alone.
4. The same decision log serves engineering debugging, security investigation and regulatory record-keeping, provided it is structured and versioned from day one.

Placeholder

> This is a seed whitepaper for reviewing the layout. Replace the text and the PDF before launch.

Executive summary

Organisations are moving from chat assistants to agents that act: they file tickets, change records, issue refunds and run code. The governance question changes with them, from *what may the model say* to *what may the system do*.

This paper proposes a reference architecture built around three responsibilities: **decide** (policy), **enforce** (a single gate in front of every tool) and **record** (a structured decision log). It then sets out a four-level maturity model and maps the architecture to the main regulatory and standards frameworks.

1. Scope and method

The architecture applies to any system in which a model can trigger side effects through tools, whether built on a vendor platform, an open-source framework or custom code. It was derived from [describe method: interviews, deployments reviewed, documentation analysed] and is deliberately technology-neutral.

2. The architecture

2.1 Decide

Policies are explicit, versioned rules over a structured request: principal (agent and the user it acts for), action (tool), resource (target) and context (amount, data classification, whether untrusted content is present). Policies live in a repository, are reviewed like code and are tested against a table of expected decisions.

2.2 Enforce

A single enforcement point sits on the only path from agent to tool. It obeys four outcomes: allow, allow with limits, escalate to a human, deny. Credentials for tools are held by the enforcement point, not by the agent, so bypassing the gate means having no credentials.

2.3 Record

Every request and decision is written to an append-only log with the policy version, the reason and, for escalations, the reviewer's identity and response. The log is the source for debugging, incident response and compliance evidence.

3. Maturity model

Level	Name	Characteristics	Typical gap
0	Implicit	Agent holds broad credentials; behaviour constrained only by the prompt	No enforcement, no record
1	Allowlisted	Fixed list of tools; basic logging	Arguments and context unchecked
2	Policy-gated	Every call evaluated by versioned policy; escalation for high-risk actions	Policies not tested; taint not tracked
3	Governed	Tested policies, per-agent identity, taint tracking, circuit breakers, audit-ready logs	Continuous evaluation and red teaming

4. Mapping to frameworks

Architecture element	EU AI Act (high-risk systems)	ISO/IEC 42001	NIST AI RMF
Versioned policies	Risk management system; technical documentation	AI risk treatment; operational planning and control	Govern; Manage

Architecture element	EU AI Act (high-risk systems)	ISO/IEC 42001	NIST AI RMF
Enforcement point	Accuracy, robustness and cybersecurity	Operational controls	Manage
Human escalation	Human oversight	Controls for human oversight	Manage
Decision log	Record-keeping (automatic logging)	Documented information; monitoring	Measure; Manage

This mapping is indicative and is not legal advice. Check obligations against the current texts and guidance.

5. Recommendations

1. Introduce a single enforcement point before adding more tools.
2. Issue each agent its own identity and remove shared human credentials.
3. Track taint from untrusted inputs and make policy depend on it.
4. Version policies and logs together so any past decision can be explained.
5. Review escalation rates monthly: too many and reviewers stop reading; too few and the thresholds are probably wrong.

Appendix: glossary

Agent. A system in which a model selects and triggers actions through tools. **Enforcement point.** The component that applies policy decisions on the path to a tool. **Taint.** A flag recording that untrusted content influenced a request.

Sources

1. Regulation (EU) 2024/1689 (Artificial Intelligence Act) . Official Journal of the European Union, 2024.

2. ISO/IEC 42001:2023 Artificial intelligence — Management system . ISO, 2023.

3. Artificial Intelligence Risk Management Framework (AI RMF 1.0) . NIST, January 2023.