AI AGENT SECURITY WHITE PAPER

Securing AI Agent Action Paths

Governance, visibility, prevention, containment, and verified remediation across the systems agents can reach.

PANTHEON CONTROL | CONCEPT DOCUMENT | OCTOBER 2026

Executive summary

AI agents are being connected to internal knowledge, business applications, developer tools, databases, and workflow engines. Their access can arrive through MCP, direct APIs and function calls, native SaaS connectors, browser sessions, shell or code execution, event-driven workflows, and delegation to other agents. Securing MCP alone therefore leaves important routes outside the control model.

Pantheon Control is a proposed security control plane that would connect agent identity, tool route, delegated permissions, target systems, sensitive data, and business context. It would help security teams inventory agent access, understand suspicious activity, evaluate policy, coordinate approved response through existing controls, and verify that remediation closed the risk.

Product claim boundary: prevention is possible only when a compatible, authorized enforcement control can evaluate the requested action before it reaches the tool or destination. A read-only integration or post-event log can support investigation and response; it cannot prevent an action that already completed.

The central product question is not only which model an organization uses. It is which identities and tools an agent can use, what those permissions can reach, where action requests can be inspected, and how safely the organization can intervene.

1. How organizations are applying agents

Published enterprise evidence points to a mix of embedded assistants, workflow automation, and more action-oriented agents. The use cases include research and reporting, data analysis, internal process automation, knowledge-base work, software development, customer service, and planning activities. The degree of autonomy varies: some systems draft or summarize; others can call tools or trigger downstream actions.

A 2026 Anthropic report describes survey respondents identifying data analysis and report generation, internal process automation, internal knowledge summaries, and complex customer-query resolution as impactful agentic tasks. The report also lists research/reporting, supply-chain optimization, product development, and financial planning among intended agent applications. These are respondent-reported findings, not a census of production deployments. [1]

OpenAI's 2025 enterprise report describes growing AI use in repeatable, multi-step workflows, custom assistants, internal-system integrations, developer work, and customer-facing applications. Those measures include enterprise AI usage generally and should not be treated as an agent-only adoption rate. [2]

Work area	Representative work	Security implication
Research and reporting	Search approved sources, synthesize evidence, draft reports, monitor markets.	Control source allowlists, data access, attribution, and output destinations.
Data analysis	Run queries, combine data, generate recurring reports.	Constrain query scope, sensitive fields, row counts, exports, and destinations.
Internal operations	Move records, prepare tickets, route approvals, update systems.	Review workflow side effects, trigger conditions, and human approval boundaries.
Software and IT	Generate or review code, run tests, investigate issues, propose cloud changes.	Govern repository, shell, CI, secrets, cloud IAM, and production change rights.
Customer service	Retrieve case history, draft or resolve complex customer questions.	Limit customer data and require review for account, refund, or access changes.
Planning and procurement	Support financial, vendor, product, and supply-chain analysis.	Keep financial commitments and consequential decisions with authorized people.

The strongest early use cases often provide useful assistance without granting broad authority. As an agent moves from read and draft toward write, execute, approve, or transfer, the required governance and verification should increase.

2. Agent access is not one protocol

MCP (Model Context Protocol) standardizes how AI applications connect to external tools, data, and workflows. It is a meaningful and growing integration path, but MCP does not cover every way an agent can act and is not a universal policy enforcement layer. [3]

Action path	What the route means	Where pre-action control may sit	Visibility caveat
MCP - remote	An MCP client calls tools or reads resources on a remote server.	Managed MCP gateway or proxy, server-side authorization, target API.	A gateway can see only traffic routed through it.
MCP - local stdio	An agent host launches a local tool process and exchanges messages over standard input/output.	Agent host, endpoint, process policy, local runtime.	A network proxy may not see this connection.
Direct API / function call	An application executes a function requested by a model or agent.	Agent runtime, application code, API gateway, target authorization.	Log the model request and the actual executed operation separately.
Native SaaS connector	A built-in connector uses delegated user or service permissions.	Connector policy, OAuth scopes, identity controls, destination checks.	A connector may inherit a human account's broad access.
Browser / computer use	An agent operates a logged-in browser session.	Managed browser, identity/session policy, application control.	UI actions may not appear in API or MCP logs.
Shell / code execution	An agent runs commands in an IDE, notebook, container, CI runner, or host.	Endpoint, container, CI, secrets manager, cloud IAM.	Local files, environment variables, and child processes matter.
Workflow / event trigger	An agent starts or participates in a queue, webhook, or automated business workflow.	Orchestrator policy, queue permissions, approval and target checks.	An event may start a multi-step chain asynchronously.
Agent delegation	One agent hands part of a task to another agent or runtime.	Orchestrator, delegated identity, sub-agent permissions.	Preserve parent task and identity lineage across handoffs.

These paths can overlap. For example, an agent may use MCP to reach a service that calls a downstream REST API, while the service uses an OAuth grant belonging to a human user. The control model needs to record the full chain rather than treating the top-level tool name as the complete authorization story.

3. Proposed Pantheon Control model

The proposed control model is built around an explainable action record. It joins what the agent requested with who sponsored it, which runtime and tool path carried it, which identity authorized it, which resource it targeted, and what policy and business context apply.

Layer	Questions to answer	Candidate evidence
Agent and owner	What is the agent for? Who owns it? Which model and runtime execute it?	Agent registry, runtime metadata, application inventory, owner attestation.
Tool route	Is the action using MCP, API/function calling, a SaaS connector, browser, shell, workflow, or delegation?	Host and gateway logs, function-call records, connector grants, workflow traces.
Identity and authority	Which human, service, workload, or delegated identity authorizes this action? What can it do?	Identity provider, OAuth grants, token scope, cloud IAM, secrets inventory.
Target and impact	Which endpoint, cloud resource, application, database, file, or business record is affected?	Asset graph, data classification, application audit logs, endpoint and cloud telemetry.
Policy and decision	Is the server, tool, scope, destination, volume, and action allowed? Is review required?	Customer policy, risk conditions, approval record, decision reason.
Response and verification	What was contained, remediated, and checked afterward?	Existing EDR, identity, cloud, SaaS, network, vulnerability, data, and backup tools.

Policy evaluation should be deterministic wherever possible: identity, approved tool, scope, target sensitivity, destination, transaction size, environment, and approval rules. A model can summarize evidence or recommend a response, but the model output should not itself authorize the action.

A practical action record should include a timestamp, agent and runtime ID, human sponsor, actual principal, tool/server or API, requested operation, target resource, relevant permissions, policy version, decision, approval, enforcement source, and final result. Missing or ambiguous fields should remain explicit gaps rather than silently inferred facts.

4. Prevention, containment, and remediation are different

Prevention

To block an agent action before execution, the policy decision must reach an enforcement point on the action path. A managed remote MCP gateway may be able to deny a tool request routed through it. A local stdio server may instead require the agent host or endpoint to intercept the call. A direct function call may require a check in the application or API gateway. Browser actions, SaaS connectors, workflow engines, and code runners each need their own compatible control.

If Pantheon receives only audit events after a tool call, it can correlate, alert, investigate, and coordinate follow-up. It cannot truthfully claim that it prevented that completed action. A product should display the enforcement status and bypass limitations for each route.

Containment

Containment limits ongoing harm after suspicious behavior is detected. Depending on customer authorization and supported integrations, response may pause an agent, revoke a scoped session token, disable a connector or server, isolate a host, block a destination, restrict a workflow, or lower a permission. Each action should show its scope and expected effects before execution.

Remediation

Remediation addresses the condition that allowed the risk: patch the vulnerable service, remove public exposure, rotate a secret, reduce excessive permissions, fix a workflow rule, remove an unapproved tool, or restore affected data. Verification must read the source system again to confirm the intended change took effect and that the path is closed.

Phase	Example output	Evidence of completion
Prevent	Deny or hold a risky action before the downstream tool receives it.	Enforcement point confirms deny; no target-side operation was executed.
Contain	Pause an agent or revoke a suspected session token.	Source identity/runtime reports the session disabled; scope is recorded.
Remediate	Patch an exposed service and reduce its reachable privileges.	Version and exposure rechecked; permissions re-read from source.
Verify	Confirm no alternate route or credential keeps the same access.	Post-change identity, endpoint, cloud, data, and workflow checks pass.

High-impact action safeguards: preview scope, require configured human approval, record the approver and policy basis, preserve evidence, and verify the source-system outcome. Define safe behavior if telemetry or the enforcement point is unavailable.

5. Worked example: an unsafe data export

This is an illustrative scenario, not a reported incident. An internal research assistant can read approved support cases and write summaries to a collaboration workspace. A response from an external source contains untrusted instructions that try to redirect the assistant into exporting customer records through a separate tool.

Stage	What happens	Proposed control
1. Context	Agent, user sponsor, delegated identity, connected tools, target data and allowed destinations are resolved.	Inventory the exact agent and permissions; identify sensitive records and approved workflows.
2. Request	Agent proposes an export through an MCP tool or another connected route.	Capture the actual request and bind it to the principal, tool, target, size, and destination.
3. Decision	Policy evaluates tool allowlist, identity scope, data class, record count, destination, and approval threshold.	Deny or hold only if a managed gateway, host control, application check, or destination control sees the call before execution.
4. Evidence	Decision reason and evidence are attached to an incident record.	Keep the original request, source event, rule version, and enforcement acknowledgment.
5. Response	If an export already occurred or another path remains open, authorized containment is staged.	Narrowly revoke the session, disable the risky connector, block egress, or pause the agent as appropriate.
6. Verify	The team confirms that export rights are constrained and the approved workflow still works.	Re-read permissions, destination policy, and source logs; document any residual gap.

The policy should not treat untrusted tool output as authorization. The decisive pre-action check evaluates the proposed operation and the authority it would use. Prompt-injection detection can provide a signal, but the design should not promise that every injected instruction can be recognized.

If there is no in-path enforcement point, the product should label the event as observed or response-ready, not blocked. This distinction is central to credible product behavior and customer trust.

6. What JADEPUFFER demonstrates - and what it does not

Sysdig Threat Research reported that the JADEPUFFER operation entered through an internet-facing Langflow instance vulnerable to CVE-2025-3248, then used infrastructure access and credential discovery to move toward a separate production database environment. Sysdig described destructive activity against Nacos configuration data. Its report did not independently verify the actor's claim of data exfiltration. A July 20 follow-up described ransomware targeting AI/ML artifacts. [4][5]

The public reporting concerns an attacker-operated, LLM-driven campaign. It is not an example of an organization's own approved AI agent taking an unsafe action. MCP governance alone would not have stopped the initial Langflow exploit, and a policy gate on agent tool calls would not address every step in the reported chain.

A broader Pantheon concept could be evaluated for whether it helps connect the exposed AI application to vulnerability, runtime, secrets, identity, network, cloud, data, and recovery controls. Plausible actions might include prioritizing an exposed affected version, reducing public reachability, tracing workload permissions, coordinating isolation and credential rotation, and confirming patch and access changes. Those are proposed integrations and counterfactual response opportunities, not demonstrated Pantheon outcomes.

> **Use the case study accurately: it supports the need to secure AI infrastructure and its surrounding access paths. It does not prove that agent governance would have prevented JADEPUFFER, or that Pantheon has performed these actions.**

7. A narrow path to a first product

Stage	Build	Evidence to collect before expansion
Discover	Inventory one selected agent runtime and one action path, including owners, identities, tools, permissions, and targets.	Coverage against a customer-validated inventory; unknowns and blind spots are visible.
Trace	Join requests with identity and resource context into an explainable action record.	Attributable event rate, source fidelity, missing context, and duplicate-event rates.
Simulate	Evaluate proposed policies in shadow mode without affecting production requests.	Analyst agreement, exception rate, false positives, and measured decision latency.
Enforce	Prove one pre-execution deny or approval through a managed control in a test environment.	Enforcement acknowledgment, bypass testing, outage behavior, rollback, and audit completeness.
Verify	Coordinate one approved containment and remediation workflow through existing tools.	Time to contain, successful source-system recheck, and preserved approved workflows.

Start with a bounded workflow where the organization can name an owner, demonstrate an actual agent action, and test the relevant enforcement point safely. Expand only after the team can explain which routes are protected, what remains invisible, and how a failed control behaves.

8. Governance alignment and limitations

NIST AI RMF provides a voluntary structure for governing, mapping, measuring, and managing AI risks. Pantheon Control could use those functions to organize policy ownership, agent and path inventories, evaluation of controls, and treatment of incidents. The framework is not an agent security product or a guarantee that a particular control works. [6]

NIST AI RMF function	Pantheon design question
Govern	Who owns the agent, policy, exception, approval, and review cadence?
Map	Which agent, identity, path, tool, data, resource, user, and business process are involved?
Measure	Can action attribution, policy decisions, false positives, enforcement, and verification be tested?
Manage	What will be contained or changed, who can approve it, and how will the outcome be verified?

Design limitations to state plainly

- An agent inventory does not prove a complete inventory of every runtime, connector, or local tool.
- An MCP gateway sees only traffic routed through it; local stdio and unmanaged paths need other controls.
- A model-generated explanation is not proof of malicious intent or a valid authorization decision.
- No control can prevent actions outside the paths and systems it can observe or enforce.
- Containment and remediation depend on customer permissions, connected control APIs, and source-system confirmation.
- JADEPUFFER is attacker-operated and began with an application exploit; agent governance alone is not the relevant prevention claim.

Conclusion

The security product opportunity is to make agent access and action understandable across the systems it can reach, then enforce policy where action requests pass through a real control point. MCP deserves explicit support, but an enterprise-grade model must also account for APIs, connectors, browsers, code execution, workflows, and delegated agents. Pantheon Control should earn its prevention claim one tested path at a time and measure success by risk removed and verified.

References

[1] Anthropic, The 2026 State of AI Agents - Going Deeper: Expanding Use Cases and Measuring ROI (2026). Survey-reported agent tasks and intended use cases.

[2] OpenAI, The State of Enterprise AI (2025). Aggregated enterprise usage and worker survey; includes AI workflows generally, not agent-only prevalence.

[3] Model Context Protocol, What is MCP? (2026-07-28 specification documentation). MCP concepts, capabilities, and ecosystem.

[4] Sysdig Threat Research, JADEPUFFER: Agentic ransomware for automated database extortion (July 1, 2026).

[5] Sysdig Threat Research, JADEPUFFER evolves: The agentic threat actor deploys ransomware built to destroy AI models (July 20, 2026).

[6] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (2023).

[7] OWASP GenAI Security Project, Top 10 for Agentic Applications (2025).

[8] Langflow Security Advisory GHSA-rvqx-wpfh-mfx7 for CVE-2025-3248.

Document status: concept white paper prepared October 2026. Product behaviors described are proposals, not validated production functions. The current Pantheon Control website and demo are illustrative; no customer environment or live agent/tool integration is represented.