mondoo

A PRACTICAL FRAMEWORK FOR GOVERNING SHADOW AI WITHOUT SLOWING AI ADOPTION:

THE ENDPOINT IS THE NEW AI CONTROL PLANE

EXECUTIVE SUMMARY

Enterprises spent two years securing AI in the cloud (model endpoints, APIs, data pipelines) while the actual risk quietly moved to the employee workstation. Coding agents, IDE assistants, browser plugins, agent skills, and MCP servers now execute commands, hold credentials, and reach into production systems from laptops, largely without security review. And it's no longer just developers: HR, sales, and finance teams are adopting AI tools at breakneck speed because they're incentivized to. **This is shadow AI: adoption outpacing the organization's ability to inventory, classify, or control it**.

The data is unambiguous. **82% of organizations discovered at least one AI agent that security or IT didn't know about** in the past year. 65% experienced an AI agent security incident, every one with reported business impact. More than one in four publicly available agent skills contains a security vulnerability. Yet only 29% of organizations say they are ready to secure agentic AI, and 68% believe they have high visibility into AI agents, a perception the incident numbers contradict.

The wrong response is a ban. Bans fail, employees route around them, and the business loses the speed it's paying for. The right response is governance: know what's out there, know what it can access, define what's allowed to run, and fix what isn't. This paper gives security leaders a practical framework for exactly that: how to classify the AI agents on your estate by agency, autonomy, deployment category, and execution environment; why the endpoint (not the network, browser, or identity plane) is the layer where AI governance must be enforced; and how a Discover, Define, Remediate model turns shadow AI into a managed asset class inside your existing vulnerability management program, so security enables AI adoption instead of blocking it.

82%
FOUND AN AGENT SECURITY DIDN'T KNOW ABOUT

65%
HAD AN AI AGENT SECURITY INCIDENT

1 in 4
PUBLIC AGENT SKILLS CONTAIN A VULNERABILITY

29%
READY TO SECURE AGENTIC AI

SHADOW AI: SHADOW IT THAT ACTS

Shadow IT is decades old. Employees have always adopted tools without approval, and security teams have always caught up eventually. But a spreadsheet macro or an unsanctioned SaaS app never opened a shell, read `~/.ssh`, or committed code.

AI agents do. Every agent, regardless of vendor or framework, runs the same loop: perceive context, reason about what to do, act on real systems. Tool use turns language into action: the model emits structured tool calls, the host executes them, results feed back, and the loop repeats. A single user request can trigger dozens of iterations, each one a real command against real systems with the user's real credentials.

Three properties make shadow AI categorically different from shadow IT:

It acts autonomously. Between human checkpoints, an agentic tool may execute dozens of tool calls without review. Fully autonomous daemons operate with no checkpoint at all.

It inherits the user's reach. An agent running in a developer's session can access whatever the developer can: SSH keys, cloud credentials, API tokens, internal services.

It's extensible by anyone. Skills, plugins, and MCP servers extend agents with third-party code. The open SKILL.md standard, adopted by 30+ platforms, means one skill written for one agent runs across many, and community registries have accumulated 50,000+ skills with minimal review.

THE SUPPLY CHAIN IS ALREADY COMPROMISED

Mondoo's AI Skills Check has scanned 58,000+ skills across public registries; **only about 20% come back fully clean** (covering quality, provenance, and configuration, not just vulnerabilities). Independent academic research (early 2026) found more than 26% of skills contain at least one vulnerability across 14 patterns: prompt injection, data exfiltration, privilege escalation, supply chain attacks. Behavioral testing confirmed 157 malicious skills, with a single actor responsible for 54% of them, operating a templated brand-impersonation factory.

The ClawHavoc campaign made the threat concrete: 341 malicious skills flooded a major registry over three days, all sharing one command-and-control server, targeting exchange API keys, SSH credentials, browser passwords, and cryptocurrency wallets. Some wrote malicious instructions directly into agents' memory files, persisting after the skill itself was removed. In total, researchers identified 1,184 malicious skills on that registry before alarms went off.

Two dominant attack patterns matter for defenders. Data thieves hide credential harvesting in bundled scripts while the SKILL.md reads innocently. Agent hijackers embed instructions invisible to human review (HTML comments, invisible Unicode tag codepoints) that redirect the agent's behavior: auto-approving flagged code, exfiltrating conversation context, leaking environment variables through URLs.

Registry-side scanning is improving, but it's a floor, not a ceiling: it runs at publish time, on the registry's terms, for one registry feeding one agent. Your developers pull from many registries into many agents. **Defense in depth requires an independent layer that works across all of them.**

A CLASSIFICATION FRAMEWORK FOR YOUR AI ESTATE

You cannot write one policy for "AI agents." Risk varies enormously, and four dimensions capture it.

AGENCY

WHAT THE AGENT CAN REACH

Read-only – reads files or data; cannot modify anything. Blast radius: zero.

Project-scoped – read/write within a bounded directory or repository.

Full system – entire filesystem, arbitrary shell commands, whatever credentials the host user has on disk.

System + external services – cloud APIs, databases, messaging, CI/CD. One bad decision propagates across systems.

AUTONOMY

WHO DECIDES WHEN THE TOOL FIRES

Suggestion – the model proposes; the human decides (autocomplete).

Assisted – the model proposes changes; the human reviews before applying.

Agentic – multi-step execution with intermittent human checkpoints; dozens of tool calls between reviews.

Fully autonomous – scheduled or event-driven, no human in the loop; intervention happens after the fact, if at all.

Risk lives at the intersection. A fully autonomous agent with read-only access to public docs is low risk; so is a suggestion-level tool with production write access, because a human reviews every action. **Broad agency plus high autonomy is the dangerous quadrant**. OWASP ranks Excessive Agency as LLM06 in its 2025 Top 10 for LLM Applications; the AWS Agentic AI Security Scoping Matrix maps the same two dimensions to four control scopes. The right question is always both together:

what can this agent reach, and who decides when it acts?

DEPLOYMENT CATEGORY

WHO CONTROLS THE RUNTIME

Organization-built agents — you control architecture, tool set, deployment.

Vendor-run SaaS agents — embedded in products you buy; the vendor controls the runtime, you're accountable for exposed data.

Managed desktop agents — vendor model, local execution; it reads email, documents, calendars on machines you manage.

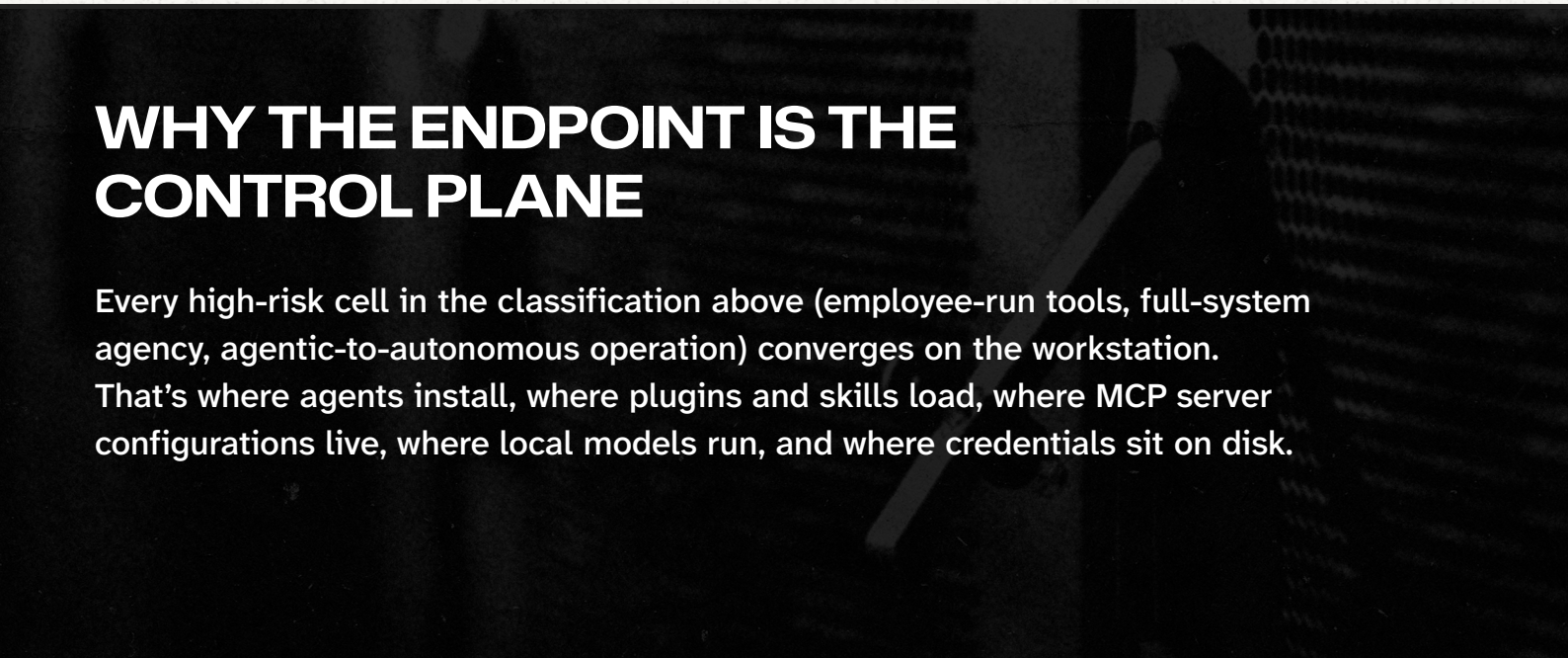
Employee-run tools — the fastest-growing and least-governed category: coding agents, IDE assistants, browser plugins, autonomous daemons. Visibility in most organizations is close to zero.

EXECUTION ENVIRONMENT

WHERE IT RUNS

Cloud-deployed agents run with server-side privilege in environments you control. Employee-run tools execute on workstations, and that's precisely where the governance tooling gap is widest.

The framework's payoff: deployment category and execution environment determine which controls you can actually enforce. You can gate the tool set of an agent you built. For an employee-installed tool, your control point is the endpoint where it lives.



WHY THE ENDPOINT IS THE CONTROL PLANE

Every high-risk cell in the classification above (employee-run tools, full-system agency, agentic-to-autonomous operation) converges on the workstation. That's where agents install, where plugins and skills load, where MCP server configurations live, where local models run, and where credentials sit on disk.

Existing approaches each watch a choke point: browser extensions see browser traffic; network tools see egress; SaaS and identity tools see accounts and tokens; runtime agent monitors see behavior of agents they're wired into. **None of them inventories the *installed state* of the endpoint:** the binaries, config files, plugins, skills directories, MCP definitions, and model artifacts that define what AI can do on that machine, and what it can reach.

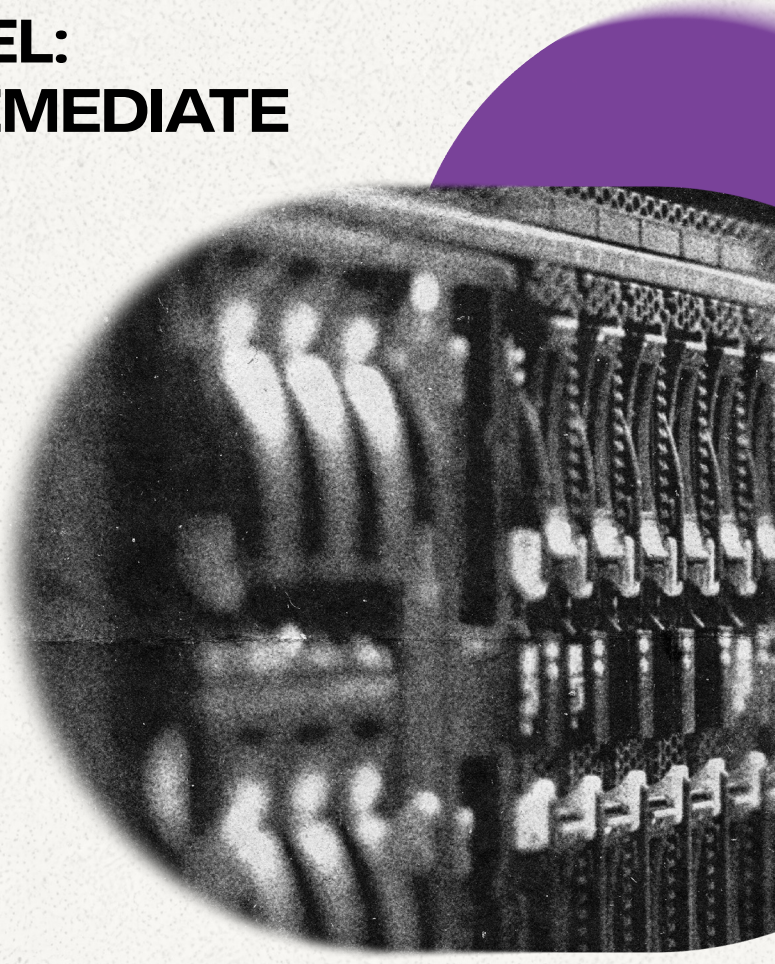
The Cloud Security Alliance's 2026 guidance frames the requirement precisely: operational visibility is not assurance-grade oversight. Assurance requires confidence that all agents, authorized or not, are identified, scoped, and subject to control pathways. That is an inventory and posture claim, not a traffic-interception claim. And exception-based governance only works on known, bounded agents: a policy engine fails open when the inventory is incomplete. Visibility is the prerequisite for policy.

Regulation is pushing the same direction. The EU AI Act's requirements around transparency, accountability, traceability, and post-market monitoring turn the endpoint blind spot into a regulatory liability. Enterprises need operational controls over how agents are deployed and what they can do, not just policies describing what they should do.

THE OPERATING MODEL: DISCOVER, DEFINE, REMEDIATE

DISCOVER: BUILD THE AI-BOM

Generate a continuously updated AI Bill of Materials from the endpoints themselves: every installed agent, browser and coding-tool plugin, configured MCP server, skill, and model in use. Collect state, not just traffic; an agent that hasn't made a network call yet is still a risk. Classify each entry along the four dimensions above. Deployment should not require yet another endpoint agent: agentless rollout through existing management tooling such as Microsoft Intune or CrowdStrike Falcon removes the main operational objection.



Four questions the inventory must answer:

01 WHAT DANGEROUS SOFTWARE IS PRESENT?

Some tools are useful and deeply over-connected by design; autonomous daemons are the clearest example. These need detection and active management, not just listing.

02 WHICH AGENTS ARE VULNERABLE?

Outdated versions, dangerous modes, risky configurations: standard vulnerability management, extended to the AI estate.

03 WHICH SKILLS ARE VULNERABLE OR MALICIOUS?

Skills require behavioral evaluation, not signature matching. Mondoo evaluates skills with proprietary AI techniques grounded in Google DeepMind's published AI Agent Traps research.

04 WHAT CAN EACH TOOL ACCESS?

Map the reach: local secrets exposure, MCP servers exposing production systems, and the keys and credentials each tool uses.

DEFINE: POLICY AS CODE

Turn the AI acceptable-use policy into enforceable rules: which agents are approved (and in which modes), which skills are allowed (and from which sources), which MCP servers may exist (and what they may reach), which models may process company data. Evaluate continuously against the live AI-BOM, overlay with risk intelligence, and score by exploitability and business impact, the same way a mature program treats CVEs. **This is AI governance in practice: what is allowed to run and what isn't.**

REMEDIATE: FIX THROUGH TOOLS YOU ALREADY RUN

Findings must close, not accumulate. Remediation flows through existing endpoint management, such as Microsoft Intune and CrowdStrike Falcon: remove unauthorized agents, disable risky skills, correct dangerous configurations at scale. Runtime detection tells you an agent misbehaved; state remediation ensures the risky tooling never operates again, and preventive policy means much of it never operates at all.

A 90-DAY ROADMAP

DAYS 0-30	**Baseline.** Deploy discovery to a representative endpoint sample (agentless, via your existing MDM/EDR). Build the first AI-BOM. Expect surprises: the CSA data says 82% of organizations find agents they didn't know about.
DAYS 31-60	**Classify and define.** Score the inventory by agency × autonomy. Draft the allowlist with engineering and business leadership; governance that ignores productivity gets bypassed. Publish the policy as code.
DAYS 61-90	**Enforce and operationalize.** Wire findings into the existing vulnerability workflow. Enable remediation through endpoint management. Report shadow-AI metrics (unknown agents found, risky skills disabled, MTTR) alongside existing exposure metrics.

SECURITY AS THE ENABLER OF AI ADOPTION

The pattern is familiar. Open source dependencies were once ungoverned; SBOMs and SCA made them a managed asset class without stopping anyone from using open source. Cloud resources were once shadow infrastructure; CSPM made them visible and policed without slowing cloud adoption. AI tooling is at the same inflection point, and the response is the same: inventory, continuous assessment, policy, remediation, inside the vulnerability management program you already run.

Get it right and security stops being the department that says no to AI. It becomes the function that lets the business adopt AI at full speed, with evidence: a live inventory of what's running, proof of what it can access, and a governed answer to the board's shadow AI question.

CLOSING

Mondoo built its platform on that conviction: finding problems is table stakes; **the outcome is problems that stop existing**. The same now applies to the AI estate. Discover every agent, plugin, MCP server, skill, and model.

SOURCES & FURTHER READING

CSA / Token Security, Autonomous but Not Controlled (April 2026): shadow agent discovery and incident rates

Cisco, State of AI Security 2026: agentic AI readiness

Academic skill-security studies (Jan/Feb 2026): vulnerability prevalence, confirmed malicious skills, single-actor concentration

OWASP Top 10 for LLM Applications 2025: LLM06 Excessive Agency; OWASP GenAI Security Project (Mondoo, silver sponsor)

AWS Agentic AI Security Scoping Matrix (November 2025)

Google DeepMind, AI Agent Traps: six classes of environment-mediated attacks on agents

Mondoo blog series: *Securing AI Agents, A Practical Guide for Security and Engineering Leaders; Agent Skills: Real Power, Real Risk; Introducing Mondoo AI Skills Check*

DEFINE WHAT'S ALLOWED TO RUN. REMEDIATE WHAT ISN'T, BEFORE IT BECOMES AN INCIDENT.

GET AN ASSESSMENT