Shadow AI Agents: How to Find the Agents Nobody Registered

Machine Identity / Last updated 2026-06-10 / https://www.scrambleid.com/learn/finding-shadow-ai-agents

Your inventory says you have a few dozen AI agents. Your identity provider's consent logs say otherwise.

TL;DR (canonical)

- A **shadow AI agent** is any AI agent, copilot, or automated tool acting on enterprise data or systems without being registered, owned, and governed by your security program.
- The exposure is measured, not hypothetical: in IBM's 2025 Cost of a Data Breach report (research by Ponemon Institute), **one in five breached organizations reported a breach involving shadow AI**, and organizations with high levels of shadow AI carried an average of **$670,000 in higher breach costs**. Only 37% had policies to manage AI or detect shadow AI at all.
- Discovery has five reliable hunting grounds: **OAuth consent grants, secrets and service-account sprawl, egress to model and MCP endpoints, SaaS-embedded agents, and procurement signals**.
- An inventory is a snapshot. The durable control is identity: when every sanctioned agent carries its own identity and signs every call, an unregistered agent has nothing to present. Discovery stops being a quarterly hunt and becomes an exception path.

Why do shadow agents exist at all?

Because adoption runs ahead of registration, every time. An MIT Project NANDA study of 52 organizations found workers at **over 90% of companies regularly using personal AI tools for work, while only 40% of those companies had purchased an official subscription**. The gap between those two numbers is the shadow estate: tools doing real work on real data with no owner of record.

Agents compound the problem. A chatbot leaks what you paste into it. An agent holds credentials, calls APIs, and acts. Gartner expects **40% of enterprise applications to ship with task-specific AI agents by the end of 2026, up from under 5% in 2025**, which means the agents arriving inside your SaaS stack will soon outnumber the ones your teams build deliberately. Most of them will arrive without a ticket.

The cost of not looking is also measured. In the IBM/Ponemon 2025 data, shadow-AI incidents exposed personally identifiable information at a higher rate than the global average (65% vs 53%), and **97% of organizations that suffered an AI-related breach lacked proper AI access controls**. The pattern auditors will recognize: the breach didn't come through the AI program. It came through the AI nobody knew was running.

Where do shadow agents hide? The five hunting grounds

1. OAuth consent grants in your identity providers

Most third-party agents enter the estate the legitimate-looking way: a user clicks "allow" and the agent gets a refresh token. Pull the consent-grant logs from your identity providers and workspace admin consoles, sort by scope breadth and grant date, and review everything with mail, file, calendar, or directory scopes that nobody can name an owner for. This is the single highest-yield query in shadow-agent discovery.

2. Secrets and service-account sprawl

Agents run on credentials, and shadow agents run on credentials nobody is watching: API keys in CI variables, service accounts created for a pilot that never ended, tokens in a vault path with no owner tag. Sweep your vaults, repos, and cloud IAM for long-lived credentials that can't be attributed to a named system and a named human. Unattributable secrets are agent fuel, whether or not an agent is holding them yet.

3. Egress to model and MCP endpoints

Your network already knows which workloads talk to model APIs and MCP servers. Egress logs filtered to the major model providers' domains, plus any self-hosted inference endpoints, give you a list of callers. Diff that list against your registered agents. Anything calling a model from a workload you can't map to a sanctioned use case is either a shadow agent or about to become one.

4. Agents embedded in your SaaS stack

The fastest-growing class doesn't run on your infrastructure at all. CRM agents, copilot builders, workflow bots: they ship inside platforms you already license, get enabled by team admins, and inherit the platform's access to your data. Audit the admin consoles of your major SaaS platforms for enabled agent and copilot features, and treat each one as an identity to govern, not a feature flag.

5. Procurement and expense signals

The MIT NANDA 90/40 gap shows up in your expense system as AI subscriptions on corporate cards that never crossed a security review. It's the least technical query on this list and it routinely surfaces tools, and the agents inside them, that every technical sweep missed.

Why an inventory isn't the fix

Run all five queries and you'll have a list. The list is stale the day you finish it, because discovery is a snapshot of an estate that changes daily. The OWASP Top 10 for Agentic Applications names the underlying risk directly: **ASI10, Rogue Agents**, the agents operating outside sanctioned inventory and oversight, and **ASI03, Identity and Privilege Abuse**, what those agents do with the leaked and inherited credentials they run on.

The durable control inverts the problem. Register every sanctioned agent with its own identity, scope its authority per action, eliminate the static secrets it could leak or share, and require a signature on every call it makes. A February 2026 NIST NCCoE concept paper on AI agent identity and authorization describes the target state: every agent action tied back to an accountable identity, auditable and non-repudiable.

In that model, the shadow agent's position changes structurally. It isn't hiding in a crowd of bearer tokens anymore; it's the only caller with nothing to present. An agent that can't mint unsigned calls can't operate off the books. Discovery stops being the control and becomes the exception path: the alert that fires when something tries to act without an identity.

That's the model ScrambleID builds: per-agent identity with zero static secrets, and per-action authority with a signature on every action. Find the shadow estate with the five queries above; end the conditions that created it with identity.

Key Takeaway

Shadow AI agents are a measured, breach-correlated exposure, and one in five breached organizations already has the scar. Hunt them in the five places they actually live: consent grants, credential sprawl, model egress, SaaS admin consoles, and expense lines. Then make the durable move: per-agent identity, scoped authority, signed actions, so the next unregistered agent isn't a discovery problem, it's a failed authentication.

FAQ

What is a shadow AI agent?

An AI agent, copilot, or automated tool that acts on enterprise data or systems without being registered, owned, and governed by your security program. The defining trait isn't malice; it's the absence of an accountable identity.

How is this different from classic shadow IT?

Shadow IT leaks data when someone moves it there. A shadow agent holds credentials and acts: it calls APIs, modifies records, and triggers workflows. The blast radius is a function of its privileges, not just its data exposure, which is why agent discovery starts in your identity providers rather than your CASB.

Can't we just block the AI tools?

Blocking handles known domains, and the MIT NANDA numbers show how that goes: usage moves to personal accounts and unmanaged paths. The organizations in the IBM/Ponemon data that fared worst weren't the ones using the most AI; they were the ones with no policy, no detection, and no access controls around the AI already in use. Govern the estate you have; don't pretend you can prevent it from existing.

Do we need an agent registry?

Yes, and it works when registration is the path of least resistance: registering an agent grants it an identity and the authority to act, and staying unregistered grants it nothing. A registry backed by policy documents intent; a registry backed by identity enforces it.

Related reading

- What is non-human identity (NHI)?
- What is AI agent identity?
- Service account replacement
- AI agent tool access playbook

References (public)

- IBM, Cost of a Data Breach Report 2025 (research by Ponemon Institute): https://www.ibm.com/reports/data-breach
- MIT Project NANDA, "The GenAI Divide: State of AI in Business 2025" (July 2025): https://mlq.ai/media/quarterly_decks/v0.1_State_of_AI_in_Business_2025_Report.pdf
- Gartner press release, "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026" (August 26, 2025): https://www.gartner.com/en/newsroom/press-releases/2025-08-26-gartner-predicts-40-percent-of-enterprise-apps-will-feature-task-specific-ai-agents-by-2026-up-from-less-than-5-percent-in-2025
- OWASP GenAI Security Project, "OWASP Top 10 for Agentic Applications" (December 2025): https://genai.owasp.org/

- NIST NCCoE, "Accelerating the Adoption of Software and Artificial Intelligence Agent Identity and Authorization" concept paper (February 2026): https://csrc.nist.gov/pubs/other/2026/02/05/accelerating-the-adoption-of-software-and-ai-agent/ipd