Securing AI Agents in the Agentic World

What the technology must do, and what the business must adopt, before agents go live

Type: Briefing **Audience:** Security, identity, risk and platform owners **Version:** 1.0 **Date:** November 2026

AI agents are a new kind of actor on the network: given a goal, acting on their own, reading data you do not control, and calling tools that touch real systems, under an identity of their own. Enabling them safely is not a model problem or a tooling problem. It is an identity and governance problem, and most organisations are not set up for it. This briefing sets out what the technology has to do and, separately, what the business has to change, so agents can be adopted without quietly becoming the largest unmanaged privileged population in the estate.

1. The shift, in three ideas

Agentic identity is a new, first-class identity type

For years identity covered the workforce, external users, and non-human identities such as service accounts and workloads. Agents are none of these. They combine the **scale and sprawl of non-human identities** with the **broad, delegated access of a human user**, and then add properties neither has: they behave non-deterministically, they can act autonomously, they can create or invoke other agents and tools, and they carry no built-in accountability for what they do. The result is a compounded risk profile: old problems (standing privilege, static credentials, orphaned access, poor visibility) at a new scale, plus new problems (discovery, ownership, delegation chains, runtime authorisation, and accountability for autonomous decisions).

Identity is the control plane

Because an agent reads data you do not control and acts on its own, the one place you can reliably see it, authorise it, govern it and stop it is its **identity**. Network, endpoint and data controls still matter, but they all now depend on identity context to make decisions. For agents, identity is not a credential to be managed; it is the control plane through which everything else is enforced.

Privilege is a spectrum, not a switch

The old model treated an identity as either privileged or not. That no longer holds. An agent that reads public web pages sits at one end; an agent that can write to a finance system sits at the other. **Controls, and the cost of applying them, should scale with where the agent sits on that spectrum, by privilege and risk, not by identity type.** This is the single most useful planning idea: it tells you where to spend.

Privilege band	Example agent	Control posture
Low	Reads public web and open data, reports back	Inventory, owner, scoped read-only identity, basic logging
Medium	Reads internal data, summarises, drafts	Least privilege, just-in-time access, egress control, behaviour logging
High	Writes to systems, moves money, changes config	Zero standing privilege, runtime authorisation per action, human approval, fast kill switch

2. The technical controls

Two layers, both required. The **identity data layer** is the static picture: who the agent is, who owns it, what it may do. The **identity runtime layer** is the live picture: what it is doing right now, and whether to allow it.

Control	What it does, and why
Discovery and inventory	Continuously find every agent, including ones created outside the sanctioned path (shadow AI), the tools and data they connect to, and the high-risk assets they can reach. You cannot govern what you cannot see, and agents can be created faster than any manual process can track.
Ownership	Every agent has a named, accountable owner. Agents have no natural authoritative source (unlike people, who anchor to HR), so without deliberate ownership they become orphans that nobody can safely change, transfer or switch off. Ownership is the precondition for every lifecycle action, including stopping one.
Least privilege	Grant each agent only the access its task needs, scoped and reviewed. Useful agents tend to ask for broad access; resisting that is the main defence against a manipulated agent doing real harm.
Zero standing privilege and just-in-time access	Agents hold no persistent elevated access. Access is granted only when needed, for as long as needed, then removed. This shrinks the blast radius of a compromised or rogue agent before any other control has to act.
Runtime authorisation	Decide each sensitive action at the moment it is attempted, using live context (who, what, where, risk) rather than a static grant. A policy decision and enforcement point sits in the path of the agent's actions, inbound and outbound.
Guardrails	Behaviour boundaries on what the agent reads and does: filtering on the prompt, on inputs it ingests, and on what it sends out, so untrusted content cannot steer it and destructive actions are blocked before they run.
Observability and intent analysis	Watch the agent's behaviour for drift between expected and actual outcomes, and for sequences heading toward an unsafe action. This is what turns "an agent did something" into "an agent is about to do something it should not", which is the signal that should trigger containment.
The kill switch	A way to stop a live agent: halt its runtime and disable its identity so it cannot re-authenticate, logged, and reversible only through approval. It acts on the agent's unique identity, which is why ownership and unique identity are load-bearing. Its sensitivity and speed should scale with the agent's place on the privilege spectrum.

THE THROUGH-LINE

Every technical control hangs off identity. Discovery finds the identity, ownership makes it accountable, least privilege and zero standing privilege limit what it holds, runtime authorisation and guardrails govern what it does, observability watches it, and the kill switch acts on it. Get the identity layer right and the rest has something to attach to. Get it wrong and none of the other controls have a reliable handle.

3. What the business must adopt

The controls above are necessary but will not land without organisational change. These are the decisions and commitments the business has to make, and they are where most agent programmes stall.

Treat agentic identity as first-class, not a credential problem

Fund and own agent identity the way you own workforce identity: with discovery, governance, monitoring and lifecycle, not as an afterthought bolted onto secrets management. Agents will become the largest privileged population in the estate; budget and ownership have to reflect that.

Establish an authoritative source and an ownership model

Decide, before agents proliferate, where the record of truth for agents lives and who is accountable for each one. Tie each agent back to a sponsoring human or team at the point of creation. No agent should exist without a named owner.

Extend joiner, mover, leaver to agents

Agent lifecycle must hook into people lifecycle. When an employee changes role or leaves, their agents must be handled deliberately: removed, transferred, or re-sponsored. An orphaned agent with live access is one of the clearest and most common risks, and it is purely a governance failure, not a technical one.

Grade by privilege and risk, and spend accordingly

Adopt the privilege-spectrum model as policy. Classify every agent by what it can reach and what it can do, and apply controls proportionate to that. This stops two failures at once: under-controlling the high-risk agents, and wasting effort wrapping heavyweight controls around trivial ones.

Assign accountability for autonomous decisions

Agreed in advance: when an agent takes an action, who answers for it? Accountability cannot be an open question discovered during an incident. It should be written into the ownership model and the approval gates for high-privilege actions.

Bring shadow AI under control

Agents can be created by anyone with the right tools, including outside sanctioned channels. The business needs a sanctioned path that is easier than the unsanctioned one, plus continuous discovery to catch what slips past, and a clear policy on what agents may and may not touch.

Integrate identity data and runtime, and make governance continuous

The static view (who the agent is, what it may do) and the live view (what it is doing) must share signal, not sit in separate tools that never speak. And because agents are adopted far faster than any audit cycle, governance has to be continuous, not a periodic review. Point-in-time assessment cannot keep pace with agent creation.

> **THE FAILURE MODE TO AVOID**
>
> The common path is to enable agents for the productivity win, treat their identities as throwaway credentials, and discover months later that nobody owns them, nobody can list them, several hold standing access to sensitive systems, and there is no way to stop one in a hurry. Every item above exists to prevent that specific outcome. It is cheaper to design in than to retrofit onto a live fleet.

4. Readiness checklist

What "ready to enable agents" looks like in practice. If you cannot tick these, you are adopting agents faster than you can govern them.

- [ ] We can produce a current, complete inventory of every agent, including ones created outside the sanctioned path.
- [ ] Every agent has a unique identity and a named, accountable owner.
- [ ] Agents hold no standing privilege; access is just-in-time and removed after use.
- [ ] Every agent is classified by privilege and risk, and its controls match that classification.
- [ ] Sensitive actions are authorised at runtime, with human approval on high-privilege actions.
- [ ] Guardrails filter what agents read and send, and block destructive actions.
- [ ] Agent behaviour is monitored for drift toward unsafe actions, and that signal is actionable.
- [ ] There is a tested kill switch that stops the runtime and disables the identity, and it has been exercised, not just documented.
- [ ] Agent lifecycle is tied to people lifecycle, so a leaver's agents are handled deliberately.
- [ ] Governance is continuous, not a periodic audit.

Vendor-neutral briefing, version 1.0, November 2026. The capabilities described map onto common identity, policy, detection and response platforms; implement with the tooling already in your estate rather than adding a separate silo.