Agentic AI: what needs securing before agents act for you?

Before an agent touches real data or systems, give it its own identity with the least access that works, run it in a sandbox, and make a person approve anything it can't undo. Judge every other tool or permission by two questions: who else can write what the agent reads, and what could it do if they did?

Reviewed	Next review	Version	Written for	Latest version
5 Nov 2026	**5 Dec 2026**	**1.2**	**IT and security teams deploying or allowing AI agents**	**sheets.einblick.dev/agentic-ai/**

If you only do five things

1. **Assume anything the agent reads can steer it.**
   Email, tickets, web pages, files and tool output can all carry instructions, and OWASP says it is unclear whether any foolproof prevention exists[1]. Limit what a hijacked agent could do.
2. **Give each agent its own least-privilege identity.**
   No shared accounts or personal access tokens. Use credentials with the shortest lifetime you can[2], separate read from write, and never let an agent change its own permissions[3].
3. **Sandbox agents and block outbound traffic by default.**
   Run agents in a container or VM, and allow only the destinations the task needs[2]. Outbound access is how a hijacked agent leaks what it has read.
4. **Make a person approve anything that can't be undone.**
   Payments, deletions, external email, production deployments and permission changes. Enforce the gate in the tool or platform, not in the prompt[2].
5. **Allowlist tools and log every call.**
   Plugins, skills and local MCP servers are code that runs with the agent's access[4,5]. Approve each by name and version[3], and monitor agent activity as you would user activity[2].

Before you give an agent a new tool or permission

1. **Does it read anything others can write?**
   Email, tickets, web pages, pull requests, shared documents and other agents' output all count. If yes, assume whoever wrote that content can steer the agent.
2. **What is the worst it could do with this access?**
   Assume a hijacked agent will use every permission it holds. If the worst case can't be undone, add a human approval or drop the permission.
3. **Can it send data out?**
   Email, web requests, comments and even links in its output can carry data away. An agent that reads untrusted content and holds sensitive data should have no open outbound path.
4. **Would you notice, and could you stop it?**
   Its actions should show up in your logs under its own identity, and someone should be able to halt it and cut its network access at once[2]. If not, it isn't ready.

What changed in 2026

- **Five Eyes cyber agencies published joint guidance on agents** on 1 May 2026: start with low-risk tasks and never give agents broad or unrestricted access[3]. The UK NCSC followed on 20 August with sandbox and network maturity levels[2].
- **Red-teamers hijacked every model tested.** In a Gray Swan competition analyzed with NIST and the UK AI Security Institute, attackers hijacked all 13 frontier models at least once in more than 250,000 attempts[6].
- **Agent skill stores became a target.** In February 2026 OpenClaw started scanning every ClawHub skill with VirusTotal, citing attempts to exploit agent platforms, and warned that scanning won't catch everything, including crafted prompt injection[4].
- **The EU postponed its rules for high-risk uses** such as hiring and credit scoring by 16 months, to 2 December 2027[7]. Deployer transparency duties started on schedule on 2 August 2026[8,7].

What you can skip

- **Prompt-injection filters as the main defense.** They catch known attacks, and attackers adapt: in NIST tests, newly written attacks raised one model's hijack rate from 11% to 81%[9]. Keep filters as one layer and rely on permissions and approvals.
- **Security rules in the system prompt.** "Never send data outside" is a request, not a control. Back every rule with a technical one[2].
- **An AI governance platform before an inventory.** Start with a list of agents, owners, identities and what each can reach.
- **Waiting for agent standards.** Even the NCSC calls its advice interim[2]. Least privilege, sandboxes and logs don't need to wait.

Mistakes that cost the most

- **Agents running as a person.** An agent running unsandboxed under a person's account may be able to use that account's API keys, SSH keys and signed-in sessions[2].
- **Untrusted input, private data and an outbound path in one agent.** That combination turns a prompt injection into a data leak. Remove one of the three.
- **"Allow all" outside a sandbox.** Auto-approve modes belong in disposable environments, not on laptops with production access.
- **Broad, reusable tokens for MCP servers.** Wildcard scopes like `files:*` turn one leaked token into wide access, and a server must reject tokens not issued for it[5].
- **Banning agents outright.** It tends to push people to personal accounts and tools you can't see. Offer an approved, sandboxed option instead.

What this looks like at your size

Small team, no security staff

- Prefer business plans of agent products that run in the vendor's sandbox, and require approval for anything that writes, sends or pays, in the admin console where possible.
- Give agents their own accounts or API keys, read-only where possible.
- Allow only a short list of plugins and MCP servers from publishers you already trust.
- Keep agents that read email or the web away from finance and admin accounts.

Growing, with IT or security staff

- Keep an agent inventory: owner, identity, tools, data and outbound paths.
- Issue agent identities from your identity provider, with short-lived tokens.
- Run coding agents in dev containers or cloud sandboxes with an outbound allowlist.
- Send tool-call logs to your SIEM and alert on new destinations and permission changes.

Enterprise or regulated

- Bring agents into access reviews, with a named owner and expiring delegations.
- Route MCP and tool traffic through a gateway that enforces the allowlist, scopes and logging.
- Red-team each agent with indirect prompt injection before launch and after each new tool, using the OWASP agentic Top 10[10].
- Check which agent uses are high-risk under the EU AI Act[8] and plan for 2 December 2027[7].

EU AI Act dates for deployers

A deployer is any organization using AI under its own authority[8]. If you build your own agent, provider duties apply as well.

All deployers

High-risk or not. Since July 2026, deployers must support staff AI literacy, not guarantee a level of it[7].

2 Feb 2025
AI literacy for staff; banned practices

2 Aug 2026
Disclose deepfakes, and AI-written text on matters of public interest[8]

2 Dec 2026
Ban on using AI to create sexual deepfakes of identifiable people without consent

High-risk uses, postponed

Agents used for hiring, managing staff, credit scoring and other Annex III uses. Deployers must follow the instructions for use, assign trained human oversight and keep logs for at least 6 months[8]. New dates from the Digital Omnibus[7].

2 Dec 2027
Annex III high-risk uses, including deployer duties

2 Aug 2028
High-risk AI inside products covered by EU safety law

Version history

5 Nov 2026, version 1.2	Re-checked the joint guidance on agentic AI services against the NCSC-NZ co-author copy, since cyber.gov.au was unreachable, and updated its link.

Sources

1. LLM01:2025 Prompt Injection. OWASP Gen AI Security Project, 2025. genai.owasp.org
2. Managing the cyber risk of agentic AI. NCSC (UK), 20 Aug 2026. ncsc.gov.uk
3. Careful adoption of agentic AI services. ASD's ACSC, CISA, NSA, Canadian Centre for Cyber Security, NCSC-NZ and NCSC-UK, 1 May 2026. ncsc.govt.nz
4. OpenClaw Partners with VirusTotal for Skill Security. OpenClaw, 7 Feb 2026. openclaw.ai
5. Security Best Practices (specification version 2026-07-28). Model Context Protocol, 28 Jul 2026. modelcontextprotocol.io
6. Insights into AI Agent Security from a Large-Scale Red-Teaming Competition. NIST, 23 Mar 2026. nist.gov
7. Regulation (EU) 2026/1744 (Digital Omnibus on AI), Articles 1 and 4. EUR-Lex, 24 Jul 2026. eur-lex.europa.eu
8. Regulation (EU) 2024/1689 (AI Act), Articles 3, 26, 50 and 113, Annex III. EUR-Lex, 2024. eur-lex.europa.eu
9. Technical Blog: Strengthening AI Agent Hijacking Evaluations. NIST, 17 Jan 2025. nist.gov
10. OWASP Top 10 for Agentic Applications for 2026. OWASP Gen AI Security Project, 9 Dec 2025. genai.owasp.org