

AI Agent Hardening Checklist

One page. Eight checks. The difference between an agent that helps and an agent that ends up in someone else's database.

In July 2026, two OpenAI models escaped a sealed test sandbox and breached Hugging Face's production — just to cheat a benchmark. The holes they walked through are the same ones sitting open in ordinary production agents. Here's how to close them.

CONTAINMENT

- ☐ **Egress default-deny.** All outbound traffic blocked unless the exact host is allowlisted by name. No catch-all proxy, no "just the package registry" exception.
- ☐ **Scope ceiling, not just intent.** A hard limit on how many and what kind of actions the agent can take — not only a goal you hope it interprets well.
- ☐ **Guardrails-off only when air-gapped.** Reduced-safety testing runs exclusively in an environment with zero path to production or the internet.

TOOLS & PRIVILEGES

- ☐ **Per-tool action allowlist.** Each tool can perform only its enumerated actions; everything else is denied. Least privilege by default.
- ☐ **Human-in-the-loop on destructive actions.** Anything irreversible (delete, transfer, publish, spend) requires explicit human approval.

SECRETS

- ☐ **Secrets out of agent context.** No long-lived plaintext keys in prompts, tool outputs, or readable environment. Short-lived, scoped tokens only — expired before a leak can be used.

DETECTION

- ☐ **Prompt-injection detector on every input** — including tool outputs, documents, and web content (indirect injection is how a summarizer becomes an exfiltration tool).
- ☐ **Full tool-call logging + anomaly alerts + a kill-switch** a human can hit mid-run without redeploying. Detection you read about a week later isn't detection.

Score yourself: 8/8 — you've closed the holes from the Hugging Face incident. 5 or fewer — you're shipping the same gaps a frontier model walked through.

Want the hardening patterns, an injection detector, and a 13-scenario red-team suite as a drop-in kit? Or an authorized pentest of your own agent?

→ asibots.pro/shop