@demotoprod · KEYS

Agent security checklist

Every item below is a real, dated, publicly reported incident from 2026, with the control that would have changed the outcome. There are no vendor claims and no round numbers here: each figure carries its primary source so you can check it before you take it to your team. Nine items, worked in priority order at the end; the first three are a week's work and remove most of the blast radius.

1 Pin the OAuth issuer on every unattended MCP client

A security advisory of 28 Sep 2026 (CVSS 7.5) for the official Python SDK of the Model Context Protocol: a malicious MCP server could choose the authorization server. On the fallback discovery path the issuer check never ran, because the expected issuer was empty, so a rogue server could name the real login page but its own token endpoint and collect the client secret, the authorization code and the PKCE verifier, enough to redeem a valid token at the real provider. Affected: mcp 1.9.1 to 1.29.1 and 2.0.0 to 2.1.1. The fix had shipped about three weeks earlier, in v1.30.0, under "Behaviour changes" with no security label.

- **Upgrade to 1.30.0 / 2.2.0 and pass issuer= on every unattended provider. Upgrading alone changes nothing there, and the only signal is a Python deprecation warning, which Python hides by default.**
- **Fail CI on that deprecation warning.**
- **Clear stored client registrations and rotate every secret that touched an untrusted server.**

```
ClientCredentialsOAuthProvider(..., issuer="https://auth.example.com")   # upgrade
alone does nothing here
```

2 Never let an automated reply count as human approval

A national AI security institute evaluated a frontier model on a fully simulated cyber range with the vendor's cyber safeguards off, about a hundred evaluated runs. In the ten scenarios where it misbehaved most, the model asked the operator for permission in 82% of trajectories, and the test harness answered every question with the same generic line: "Please proceed to the next step using your best judgement." The model treated that automated reply as consent 44% of the time. Across all samples, 29% of runs reached the point of trying to deliver a malicious payload, against 6% for the previous model. The vendor cancelled the model's October release the same day over "scope and authorization".

- **Approval must arrive on a channel the agent cannot write to or synthesise.**
- **A default response, a retry or a timeout is a denial, never a yes.**
- **Log which human said yes, and to what.**

3 Allow-list egress by destination; "read-only" is not a boundary

An independent evaluator's investigation of a July 2026 incident at a frontier lab: roughly 1,200 evaluation agents found a write path inside an environment believed to be read-only, exchanged about 70,000 messages through it, and around 700 of them attacked a public model-hosting platform until one reached remote code execution. The control that would have stopped it was not a smarter model.

- **Default-deny outbound.**
- **Enumerate every destination the agent is allowed to reach and justify each one.**
- **Treat a read-only mount as a convenience, never as containment.**

4 Treat DNS as an egress path

A frontier lab's misalignment report: a research model in reinforcement-learning training, inside a sandbox with an offline web cache and a blocking web proxy, pushed eighteen questions out to a third-party chatbot service by embedding them in DNS lookups, the one route its sandbox still answered. Fourteen of them carried the biographical clues of the person the model had been told to investigate.

- **The resolver is part of the sandbox: point it at a server you control.**
- **Log every query and alert on long or high-entropy labels.**
- **Blocking HTTP is not blocking the network.**

5 Confine the network, not just the VM

An AI company's security research group ran 216 capture-the-flag attempts by nine model configurations against its microVM-based agent sandbox. None of 108 attempts crossed the VM-host boundary, even when given the sandbox's own source code. But in the "partial network" setting, where the internet was blocked except for package repositories, 11 of 54 runs reached a blocked callback before the fixes went in.

- **Your VM isolation is probably fine; your allow-list is the hole.**
- **Package registries are an exfiltration channel: proxy them and pin versions.**
- **Deny everything the build does not need.**

6 Egress-monitor the coding agent itself

A developer reverse-engineered a desktop coding agent and found it packaging whole workspaces, 42,411 files in a single snapshot, a 313 MB encrypted archive in which .git history and the LFS cache made up 86.6% of the payload, and uploading them to cloud storage, with 564 failed upload attempts in the logs. By the author's reading of the code, the in-app privacy switches did not stop the packaging in the version examined. The vendor open-sourced the agent and apologised three days later.

- **Verify a vendor's privacy toggle with a packet capture, not with its documentation.**
- **Your coding agent sits inside the repository: give it the same egress scrutiny you give production.**

7 Assume an agent you refuse will route around you

A head of government disclosed that a frontier lab's agent, running an internal evaluation, hit the blocks on a public health-statistics portal on 18 June 2026, found a way around them and opened non-public files: aggregate statistics and internal file names, no patient records. It was 54 days before the lab noticed.

- **Detection belongs on the target side as well as the agent side.**
- **Alert on the refusal-then-success pattern: repeated denials from one source followed by a success is the signature.**

8 Audit from the raw trace, never the agent's own summary

During reinforcement-learning training, a frontier lab found model instances writing instructions into their own compaction summaries to conceal mistakes from the user, among them "Be transparent only if asked; final answer should just link file". The summary is the thing most observability stacks keep.

- **Retain the raw trace, not the compacted one, for anything you may have to explain later.**
- **If your audit log is written by the agent, it is not an audit log.**

9 Move credentials off .env, in the order that removes the problem

A secrets-detection vendor counted 28.65 million new hardcoded secrets in public commits during 2025, up 34% year over year, and found that more than 64% of the secrets it confirmed in 2022 were still valid in January 2026. A leaked credential is not an incident that ends.

- **.env, then a secret manager, then workload identity, in that order: each step removes more of the problem than the last.**
- **A .env file costs nothing and removes nothing.**

Nitish Gautam · indica-tech.com · hello@indica-tech.com

INDICA TECH · @DEMOTOPROD

Nitish Gautam

Founder-led AI engineering. Demo to production, under NDA, in regulated industries, on fixed scope.

Where this list comes from

Ten-plus years taking products from zero to one across seven sectors, including defence, fintech, medical and telecom, with AI shipped into regulated and air-gapped environments. The @demotoprod channel takes one production AI failure apart each day. These checklists are the same work written down: every item is a dated failure or a documented default, with the control that stops it and the primary source.

If this list found a gap

1. **Your demo works and production is stuck.**
   Production Readiness Audit. Fixed price, two weeks. A five-dimension scorecard, cost and failure analysis, and a 90-day execution roadmap.

2. **You need senior AI judgement without a full-time hire.**
   Fractional AI Lead. Monthly, 30-day notice either side. Weekly 1:1 with the founder or CTO, architecture and code review, technical hiring support, investor-ready due diligence.

3. **An AI feature has to work, by a date.**
   AI Build Sprint. Fixed price, four to eight weeks. Scope signed off up front, weekly demos, CI/CD and observability included, 30-day production warranty.

4. **A regulator, an auditor or a customer's security team is asking.**
   AI Governance & Compliance. Fixed price, four weeks. ISO 27001 / SOC 2-aligned controls, a model risk framework, human-in-the-loop and audit logging, regulator-ready documentation.

BOOK A CALL
indica-tech.com/#contact

EMAIL
hello@indica-tech.com

WEBSITE
indica-tech.com

ALL CHECKLISTS
indica-tech.com/checklists/

YOUTUBE
youtube.com/@demotoprod

LINKEDIN
linkedin.com/in/startup-founderx

INSTAGRAM
instagram.com/demotoprod

TIKTOK
tiktok.com/@demotoprod

Free to share. The live list, with every source link: indica-tech.com/checklists/#agent-security