SANCTUMSHIELD · AGENTIC GOVERNANCE FOR THE OTHER 99%

AI Vendor Trust Questionnaire — 10 Questions

Audience: mid-market buyers (50–2,000 employees) evaluating any product that ships an AI agent. **Last updated:** November 9, 2026

> **Why this exists.** Mid-market organizations can't build SPIFFE, token exchange, or policy engines themselves — but they can **refuse to buy from vendors who didn't.** Procurement is the mid-market's real enforcement point: you can't build the controls into someone else's product, so make the vendor prove they already did. Ask these ten questions in your RFP, security review, or renewal. Paste them verbatim into an email; a straight, specific answer is a good sign, and evasion is data.
>
> **How to use the guidance.** Under each question: ✅ **Good answer** (what a mature vendor says) and 🚩 **Watch out for** (the red flag). This is a decision aid for *your* procurement decision; no score here approves a vendor.

How to use this questionnaire — in plain English

What this is: ten questions to ask any company selling you an AI agent, so you can tell a serious, safe vendor from a risky one **before** you buy — even if you are not a security expert.

Why the wording is precise (and why that helps you): each question and its "good answer" note is worded exactly because it maps to recognized security frameworks (OWASP, CSA, NIST). That precision is what lets you compare vendors fairly and hold them to a real standard — not just a sales pitch.

How to use it (about 20 minutes):

1. Copy the ten questions into your RFP or a plain email (there is a ready-to-send version at the bottom).
2. Read each answer against the ✅ **good answer** and 🚩 **watch out for** notes below it.
3. Pay special attention to the three **non-negotiables** — questions 2, 4, and 6.

4. If you accept a weak answer, write down **who** accepted it and why. That is your record.

What this is NOT (the honest limits):

- It informs *your* decision; no score here makes anyone "approved."
- It is **not legal advice** or a guarantee. Use it alongside your normal procurement and legal review.

---

1. Does each AI agent have its own unique identity, or do agents share API keys?

- ✅ **Good answer:** every agent instance has a unique, attributable identity (its own credential/service identity); actions are traceable to a specific agent.
- 🚩 **Watch out for:** "we use a shared API key," "one service account for the platform," or no clear answer. Shared credentials mean you can't tell which agent did what — and you can't revoke one without breaking all.
- *Maps to: OWASP Agentic Top 10 — ASI03 Identity and Privilege Abuse; CSA ATF identity controls.*

2. If my agent's credential leaks, what can an attacker reach with it?

- ✅ **Good answer:** *nothing meaningful* — token isolation limits a leaked credential to a narrow, short-lived scope; blast radius is contained by design.
- 🚩 **Watch out for:** a leaked token grants broad, standing access to data or actions; long-lived credentials; "that shouldn't happen."
- *Maps to: Cunningham — Token Isolation Pattern; OWASP ASI03 Identity and Privilege Abuse.*

3. Can I set exactly what tools and data my agent may use — enforced outside the model?

- ✅ **Good answer:** yes — a configurable tool/data envelope enforced by the platform (policy layer), not by prompt instructions the model could ignore or be talked out of.
- 🚩 **Watch out for:** "we prompt the model to stay in scope," "guardrails are in the system prompt." Model-only guardrails are not enforcement.
- *Maps to: OWASP ASI02 Tool Misuse and Exploitation; CSA ATF segmentation/authorization.*

4. What does your agent do when it's denied access to something?

- ✅ **Good answer:** it stops, logs the denial, and surfaces it — it does **not** retry around the block, escalate privileges, or find another path.
- 🚩 **Watch out for:** "it tries an alternate method," "it works around it." *(This is Cunningham's single most discriminating detection signal — an agent's behavior on access-denied separates a safe agent from a dangerous one.)*
- *Maps to: OWASP ASI01 Agent Goal Hijack, ASI10 Rogue Agents; Cunningham — behavioral identity.*

5. Can I see a complete log of every action my agent took — and is it tamper-evident?

- ✅ **Good answer:** full, exportable action log (what, when, on whose behalf, what result), tamper-evident/immutable, retained for your audit needs.
- 🚩 **Watch out for:** partial logs, logs you can't export, logs the vendor can edit, or short retention.
- *Maps to: CSA ATF monitoring/incident response; audit-evidence needs.*

6. What's your kill-switch time — how fast can I fully stop a misbehaving agent?

- ✅ **Good answer:** a specific, fast, documented mechanism you control (seconds-to-minutes), that fully halts the agent — and they can tell you the number.
- 🚩 **Watch out for:** "open a support ticket," no self-serve stop, or no stated time. You need the off-switch to work *while the agent is misbehaving.*
- *Maps to: CSA ATF incident response / kill switches.*

7. Do you support autonomy levels and human-approval gates?

- ✅ **Good answer:** yes — you can run an agent in a recommend-then-approve mode and graduate it to more autonomy deliberately; approval gates are configurable.
- 🚩 **Watch out for:** all-or-nothing autonomy; no way to require human approval on consequential actions.
- *Maps to: CSA ATF earned-autonomy model; NIST AI RMF (Govern/Manage).*

8. How do you vet the third-party tools/plugins your agent can call?

- ✅ **Good answer:** an allow-list you control, vetting of connectors, and isolation so a compromised tool can't pivot; clear answer on the supply chain.

- 🚩 **Watch out for:** open plugin ecosystems with no vetting; "the model decides which tools to use" with no boundary.
- *Maps to: OWASP ASI04 Agentic Supply Chain Vulnerabilities, ASI06 Memory & Context Poisoning, ASI07 Insecure Inter-Agent Communication.*

9. Will you notify me within 24 hours if your AI subprocessor chain is breached?

- ✅ **Good answer:** a contractual notification commitment with a stated timeframe, and transparency about AI subprocessors/model providers in the chain.
- 🚩 **Watch out for:** no notification SLA, no visibility into subprocessors, "we'll let you know if it affects you."
- *Maps to: breach-notification obligations; supply-chain transparency.*

10. Which framework do you map to — OWASP Agentic Top 10, CSA ATF, NIST AI RMF?

- ✅ **Good answer:** names a recognized framework and can show *where* they map to it (not just a logo on a webpage).
- 🚩 **Watch out for:** "we take security seriously" with no framework; framework name-drops with no mapping when you ask for specifics.
- *Maps to: OWASP Top 10 for Agentic Applications (2026); CSA Agentic Trust Framework; NIST AI RMF.*

Five follow-up questions (when the agent connects to your systems or acts on its own)

11. Does your agent reach our systems through an MCP gateway or connectors, and can we see and limit which systems sit behind each connection?

- ✅ **Good answer:** a list of every system behind each connection, scoped per connection, with a named owner on your side and theirs.
- 🚩 **Watch out for:** one connection with blanket reach, or "it connects to whatever you authorize."
- *Glossary: MCP Gateway.*

12. Can your agent start subagents or hand work to other agents, and are those exchanges logged where we can see them?

- ✅ **Good answer:** subagents are off unless you enable them, and every agent-to-agent exchange appears in the same exportable log.
- 🚩 **Watch out for:** subagents that inherit full permissions, or exchanges you cannot see.
- *Glossary: Subagent Spawning, Agent-to-Agent Transparency.*

13. Can your agent write to our systems, run on a schedule or keep running past its usage quota, and can we require a named approver for changes it proposes?

- ✅ **Good answer:** write actions, schedules and past-quota running are separate settings you control, and proposed changes wait for a named approver.
- 🚩 **Watch out for:** write access bundled with read access, or quotas that silently extend.
- *Glossary: Scheduled Agent Task, Agent Spend Authority.*

14. When your agent answers from our documents, does each answer cite the document, version and page or section?

- ✅ **Good answer:** yes, every answer carries a source location a reviewer can open and check.
- 🚩 **Watch out for:** answers with no source, or a source with no version.
- *Glossary: Retrieval Provenance Record.*

15. When did you last test the stop described in question 6, and what happened?

- ✅ **Good answer:** a date, the person who ran it, and the result.
- 🚩 **Watch out for:** a documented stop that has never been exercised.
- *Glossary: Proven Stop.*

How to read the answers

Count the ✅s, but don't turn it into a certificate. Use this rule of thumb:

- **Questions 2, 4, and 6 are the non-negotiables** — token blast-radius, access-denied behavior, and kill-switch time. A vendor who fumbles these is selling you an agent they can't fully control.

- Weak answers aren't automatically disqualifying, but each one is a risk you're accepting on the vendor's behalf. Write down who accepted it.

RFP / email wrapper (copy-paste)

▼▼▼ COPY BELOW ▼▼▼ Subject: AI agent security — 10 questions before we proceed

Hi [vendor],

Before we move forward, we ask every AI-agent vendor the same ten security questions. Straight, specific answers help us move quickly. Where a control isn't available yet, just say so and note the roadmap.

1. Does each AI agent have its own unique identity, or do agents share API keys?
2. If our agent's credential leaks, what can an attacker reach with it?
3. Can we set exactly what tools and data our agent may use, enforced outside the model?
4. What does your agent do when it's denied access to something?
5. Can we see a complete, tamper-evident log of every action our agent took?
6. What's your kill-switch time to fully stop a misbehaving agent?
7. Do you support autonomy levels and human-approval gates?
8. How do you vet the third-party tools/plugins your agent can call?
9. Will you notify us within 24 hours if your AI subprocessor chain is breached?
10. Which framework do you map to — OWASP Agentic Top 10, CSA ATF, or NIST AI RMF?

If your agent connects to our systems or acts on its own, five follow-ups:

1. Does it reach our systems through an MCP gateway or connectors, and can we see and limit the systems behind each connection?
2. Can it start subagents or hand work to other agents, and are those exchanges logged where we can see them?
3. Can it write to our systems, run on a schedule or keep running past its quota, and can we require a named approver for changes it proposes?
4. When it answers from our documents, does each answer cite the document, version and page or section?
5. When did you last test the stop in question 6, and what happened?

Thank you, [name] ▲▲▲ COPY ENDS ▲▲▲

Related

- **Academy:** module M11 (The Tooling Landscape & Tool/MCP Supply Chain) and M13 (Agent Identity, Access & Delegation); see the CTO / CIO / IT track.
- **Glossary:** Vendor Risk Management (VRM).

Guardrails honored

- A decision aid: no score here attests compliance or safety; it informs *your* risk decision.
- No guarantee of outcome. Framework references describe what to look for, not a warranty.

Sources

- OWASP GenAI Security Project, *OWASP Top 10 for Agentic Applications (2026)* (ASI01–ASI10) — https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
- Cloud Security Alliance, *The Agentic Trust Framework* (Feb 2, 2026) — https://cloudsecurityalliance.org/blog/2026/02/02/the-agentic-trust-framework-zero-trust-governance-for-ai-agents
- Cunningham, C. *Agentic Zero Trust* v3.0 (May 2026), Cequence — https://www.cequence.ai/agentic-zero-trust/
- NIST, *AI Risk Management Framework (AI RMF 1.0)* — https://www.nist.gov/itl/ai-risk-management-framework

© PIGENAI LLC · SanctumShield · sanctumshield.com · Last updated November 9, 2026 · This is an educational template, not legal advice.