

AI Agent Security Checklist

Fourteen checks before an agent with tools is allowed on a live operation. Aligned in spirit with the OWASP Top 10 for LLM Applications, without pretending a checklist is a certification.

UPDATED 2026-10-22

Use this before a pilot touches a real queue, a claim file, a core system, or a customer. It is a working list for a CISO, a delivery lead, and the person who owns the operation. It is not a score you can publish, and it is not a claim that S·CORP or your firm holds a certification.

The public frame is the OWASP Top 10 for LLM Applications: prompt injection, sensitive information disclosure, supply chain, excessive agency, and improper output handling are the items this list is built to make concrete. NIST's AI Risk Management Framework (Govern, Map, Measure, Manage) is a reasonable way to file the answers. ISO 27001 and PCI DSS matter only where those regimes already apply to the system.

How to use it

Answer each line for one agent, not for "AI" in general. A "no" is a reason to narrow the pilot, not a reason to hide the line. Keep the completed list with the change record for the agent.

Checks

- ☐ The operation the agent may perform is written in one sentence, with an owner.
- ☐ The actions the agent must not take are listed: money movement, limit changes, closing a case, sending data outside the firm.
- ☐ Every tool is named, with a verb (read, draft, create) and an allow-list of parameters. Free text from the model is not a parameter.
- ☐ The tool identity is not a shared administrator account. A person can see which agent called which system.
- ☐ Untrusted content (email, PDF, ticket, web page) cannot by itself change a tool parameter. A hostile instruction in a document was tried.
- ☐ The retrieval index matches a data map. Fields that are out of scope, including cardholder data, are not embedded.

- ☐ Secrets, internal host names, and fraud-rule logic are not in the system prompt. Assume the prompt can be quoted back.
- ☐ Model output is treated as untrusted before it is shown to a customer or written to a system of record.
- ☐ A person approves any action that changes a record. The approver sees the tool name and the parameters, not only a summary.
- ☐ A named person can take the agent off the live path the same day, without a code change, and the stop is recorded.
- ☐ Each unit of work records the case id, tools called, data classes touched, output shown, and the human decision.
- ☐ The contract says where prompts are stored, who can read them, how long traces live, and that customer content is not used to change a model unless you agreed.
- ☐ Retention of prompts, embeddings, and traces is no longer than the underlying operation, and deletion has an owner.
- ☐ Someone samples transcripts against the written scope on a schedule the operator recognizes. The sample method is written down.

What this list will not do

- It will not tell you which model to buy.
- It will not classify your system under the EU AI Act. That is a legal judgment.
- It will not replace a penetration test of the surrounding applications.
- It is not a case study and it does not cite results from a client.

S-CORP — Engineered for outcomes. Built for trust.

s-corp.app · AI Agent Security Checklist