AI Agent Identity Readiness Checklist

Five dimensions to verify before any AI agent enters production.

> AI agents are not just another software component. They have identity, capabilities, and access patterns that existing security programs were not designed to evaluate. This checklist covers the five dimensions every organization needs to assess before deploying AI agents into production. Each dimension has five specific controls. If you cannot check it, you have found work to do.

1 Governance and Policy

Every AI agent needs a defined owner, a clear scope, and documented boundaries. Without governance, agents operate in the blind spots between teams.

☐ AI usage policy exists and is current A documented policy defines what AI agents may do, what data they may access, and what outcomes are prohibited. The policy is reviewed at least quarterly.

WHY *Without a policy, teams build agents based on assumptions. Assumptions are not controls.*

☐ Agent registry maintained Every agent in production is registered with its purpose, owner, access scope, data classification, and risk level. Unregistered agents are treated as unauthorized.

WHY *You cannot govern what you do not track. Shadow AI agents are the most common source of uncontrolled data exposure.*

☐ Change management process defined Any modification to an agent's prompts, tools, data sources, or access permissions triggers a documented review process.

WHY *Prompt and tool changes can silently alter agent behavior and expand its attack surface without any code deployment.*

☐ Ethical review completed The agent's intended use case, target population, and potential for harm have been documented and reviewed by the appropriate governance body.

WHY *EU AI Act deployer obligations require documented risk assessments for high-risk systems. Even non-high-risk agents benefit from ethical review.*

☐ Decommission criteria defined The conditions under which an agent is taken offline, including trigger events, responsible parties, and data handling procedures.

WHY *Agents that accumulate capabilities without a clear path to retirement become persistent risk vectors with no owner.*

2 Risk and Compliance Mapping

AI agents introduce risk categories that do not fit neatly into existing frameworks. Map them explicitly before deployment.

☐ OWASP Agentic Top 10 mapped Each of the ten agentic-specific risk categories has been evaluated against the agent's architecture with a documented finding.

WHY *Generic application security frameworks miss agent-specific risks like tool poisoning, meta-context injection, and autonomous privilege escalation.*

☐ Data classification applied Every data source the agent accesses has been classified, and the classification determines the agent's access scope and monitoring requirements.

WHY *An agent accessing classified data without appropriate controls is a compliance violation waiting to happen.*

☐ Regulatory obligations identified All applicable regulations (EU AI Act, sector-specific rules, data protection laws) have been mapped to the agent's behavior and data flows.

WHY *Regulatory obligations for AI systems are evolving rapidly. Untested assumptions about compliance are not compliance.*

☐ Third-party dependencies assessed Every model, API, tool, and data source used by the agent has been evaluated for security posture, availability SLAs, and data handling commitments.

WHY *A third-party model that changes its behavior or terms of service can break your agent's security controls without warning.*

☐ Incident classification defined AI-specific incidents have their own classification criteria separate from traditional security incidents, with different escalation paths.

WHY *An agent exfiltrating data through its outputs is a different class of incident than a compromised service account. Treat it accordingly.*

3 Organizational Capability and Skills

The people who build, operate, and respond to AI agents need skills that most security programs do not currently develop.

☐ AI security skills inventory completed The team's capabilities in prompt engineering security, LLM architecture, agent framework internals, and adversarial testing have been assessed.

WHY *You cannot close skill gaps you have not identified. Most teams overestimate their AI security capabilities because they have not been tested.*

☐ Training program in place Developers, security engineers, and operations staff have role-specific training on agentic AI security patterns and failure modes.

WHY *Training on traditional application security does not cover prompt injection, tool poisoning, or agent privilege escalation.*

☐ Red team capability established Either an internal team or an external provider with proven agentic AI testing capabilities is available for independent validation.

WHY *Self-testing creates blind spots. An external perspective is essential for finding the failure modes your team's assumptions prevent you from seeing.*

☐ Security champion designated A named individual on the agent's team has deep expertise in AI security and serves as the first point of contact for security questions.

WHY *Security questions that go unanswered are the most common path to insecure deployments. A designated champion prevents unanswered questions.*

☐ Incident response drill conducted The team has practiced responding to at least one AI-specific incident scenario within the last six months.

WHY *Incident response plans that have never been tested are theoretical documents. Practice reveals gaps that planning alone cannot find.*

4 Procurement and Vendor Management

When you procure AI capabilities, you are acquiring not just technology but behavior, data access patterns, and risk profiles.

☐ Vendor AI security questionnaire completed The vendor's security controls for their AI systems have been documented, including their approach to prompt injection defense, data handling, and model integrity.

WHY *A vendor's AI security posture directly affects your agent's security. You are responsible for their controls once integrated.*

☐ Data handling agreement signed The contract specifies how the vendor handles your data, whether it is used for model training, retention periods, and deletion procedures.

WHY *Data submitted to a third-party AI service may be retained, used for training, or shared with other customers. Without explicit contractual controls, you have no guarantee.*

☐ SLA covers AI-specific failures The service level agreement includes metrics for AI-specific failure modes (unexpected behavior, prompt injection susceptibility, tool misuse) in addition to uptime and performance.

WHY *Standard SLAs measure availability and response time. They do not measure whether the AI system is behaving as intended.*

☐ Exit strategy documented The process for replacing the vendor's AI capability, including data portability, model migration, and service continuity, is defined.

WHY *Vendor lock-in in AI systems is harder to escape than traditional software. Planning for exit before you need it is the only time it is feasible.*

☐ Security audit rights reserved The contract grants the right to audit the vendor's AI security controls, or an equivalent third-party certification has been verified.

WHY *Vendor claims about security controls are not substitutes for verification. Audit rights create accountability.*

5 Operational Controls and Monitoring

An agent in production needs the same operational controls as any other system, plus AI-specific monitoring for probabilistic behavior.

☐ Agent identity and scope enforced The agent's credentials, tool access, and data permissions are scoped to its documented purpose and no broader. Regular access reviews are conducted.

WHY *Agents that accumulate access over time create persistent privilege escalation risks. Scope enforcement is an ongoing control, not a one-time setup.*

☐ Behavioral monitoring active The agent's outputs and tool calls are monitored for deviations from expected behavior patterns, with alerting configured for anomalous activity.

WHY *AI agents behave probabilistically. Monitoring for behavioral anomalies catches compromise and misconfiguration that rule-based detection misses.*

☐ Audit logging complete Every agent interaction (input, output, tool call, decision point) is logged with timestamps, context identifiers, and outcomes in a tamper-resistant store.

WHY *Without complete audit logs, incident investigation is impossible. Partial logs create false confidence in your ability to investigate.*

☐ Rate limiting and quota controls applied The agent's tool calls and API interactions are rate-limited and quota-controlled to bound the blast radius of a compromised or misbehaving agent.

WHY *An agent under attack or malfunctioning can generate thousands of tool calls per minute. Without rate limits, the blast radius is unbounded.*

☐ Kill switch tested A mechanism exists to immediately disable the agent's tool access and stop its execution, and it has been tested in a non-production environment.

WHY *The first time you need a kill switch should not be during an active incident. Test it before you need it.*
