ENTERPRISE READINESS BEST PRACTICES

Agentic AI Deployment Checklist

Ten security domains to evaluate before deploying AI agents. Hover any item for context on why it matters. Click to confirm it's in place.

aembit

DIRECT ATTACK VECTOR OR COMPLIANCE FAILURE HIGH
CONTROL GAP WITH MEANINGFUL EXPOSURE MED
GOVERNANCE & OPERATIONAL HYGIENE LOW

01 Data Exposure

- [ ] **Inventory of all data sources the agent can access is documented and current** — HIGH
  Agents operate on whatever they can reach. Without a current inventory, you cannot scope risk, enforce least-privilege, or respond effectively to an incident.
- [ ] **Internal systems including email, documents, and financial data are accounted for in that inventory** — MED
  Internal systems carry the highest sensitivity and the greatest compliance exposure. An agent with implicit access to productivity infrastructure, document stores, or financial data platforms is a material risk if that access isn't explicitly scoped and governed.
- [ ] **The ability to combine data across sources in ways individual users normally cannot has been assessed** — MED
  Agents naturally aggregate across data sources in ways individual users typically do not. A single agent session can combine HR data, financial records, and communications – creating exposure that no individual data source policy is designed to address.

02 Actions & Capabilities

- [ ] **Permitted action scope is explicitly defined and scoped independently of the user's permissions** — HIGH
  Read access and write access are categorically different risk profiles. An agent that inherits a user's full permissions can send email, update records, or trigger workflows — causing irreversible harm if misconfigured or compromised.
- [ ] **Behavior has been tested against edge cases including unexpected inputs, ambiguous instructions, and boundary conditions** — LOW
  Defining action scope on paper is not the same as verifying it in practice. AI agents can interpret ambiguous instructions in unexpected ways, and those gaps are easier to find in testing than in production.
- [ ] **A review process exists to adjust the agent's action limits as its use expands** — LOW
  Agent scope tends to grow incrementally and informally. Without a review process, what started as a read-only research tool quietly becomes something that can modify records or initiate communications.

03 Identity

- [ ] **The agent has a distinct identity, separate from the employee identity it operates on behalf of** — HIGH
  Without a distinct agent identity, every action an agent takes is indistinguishable from a human action in your logs. Attribution becomes impossible, and your audit trail — required for SOX, PCI, and most internal security standards — breaks down.
- [ ] **Audit logs can definitively attribute actions to the agent versus to the user directly** — HIGH
  When something goes wrong, the first question is whether a human or an agent did it. If logs cannot answer that definitively, incident response and compliance reviews both stall.
- [ ] **Identity is consistent and verifiable across all connected tools and MCP servers** — MED
  Inconsistent identity across MCP servers means access policy cannot be applied uniformly. If the agent presents a different identity to different services, enforcement gaps exist at every boundary where identity is not verified.

04 Access Control

- [ ] **Permissions are scoped per task or use case rather than derived from the user's existing access rights** — HIGH
  Permissions inherited wholesale from the user give the agent the same access rights as the person it operates on behalf of. Least-privilege requires defining what the agent needs explicitly, scoped to the task.
- [ ] **A centralized policy governs what the agent can do within each connected system** — HIGH
  Per-system, per-user configuration doesn't scale. A single policy control plane that enforces access decisions consistently across every connection is what makes agent deployments manageable at scale.
- [ ] **Access can be adjusted at a granular level without disrupting user workflows** — MED
  If tightening an agent's access requires touching every user's configuration individually, it won't happen quickly enough to respond to incidents or policy changes.

05 Credentials

- [ ] **Secrets, such as API keys and tokens, are not stored in agent configurations or MCP server config files** — HIGH
  Static secrets in configuration files are a well-documented attack vector. They aren't rotated, are frequently committed to version control, and grant access to anyone who can read the file.
- [ ] **Credentials are short-lived and scoped to specific sessions, not long-lived static secrets** — HIGH
  Long-lived credentials mean a compromised token provides persistent access. Short-lived, session-scoped credentials limit the damage window to minutes rather than months.
- [ ] **A defined rotation and revocation process exists for all agent credentials** — MED
  Without a rotation policy, credentials age indefinitely. Without a revocation path, response to a compromise is delayed. Both are findings in any regulated environment audit.

06 Audit & Visibility

- [ ] **All agent actions are logged with sufficient context to fully reconstruct what happened** — HIGH
  A log entry that says, for example, 'resource accessed' is not an audit trail. You need the agent identity, the specific tool invoked, the parameters passed, the policy decision, and the timestamp.
- [ ] **Agent-initiated activity is clearly separated from user-initiated activity in all logs** — HIGH
  Mixed logs make investigation impractical and compliance attestation unreliable. Regulators and auditors require clear attribution between agent-initiated and user-initiated activity.
- [ ] **Activity logs are mapped to your compliance framework and a complete audit record can be produced on demand** — MED
  'We have logs' and 'we can produce an audit-ready record' are different things. Controls that exist but aren't tied to your compliance framework (SOX, PCI, HIPAA, SOC 2) don't count during a review. If fulfilling a regulator or legal request requires manual log assembly across systems, that gap will surface during a review

07 Control & Incident Response

- [ ] **Agent access can be disabled immediately without affecting the underlying user's access** — HIGH
  If revoking agent access requires revoking the user's access, the two are too tightly coupled to respond effectively. Clean separation between agent access and human access is a prerequisite for incident response.
- [ ] **The incident response plan covers AI agent scenarios — containment, investigation, and notification steps are defined** — HIGH
  Most IR plans were written before agentic AI existed. An AI agent that exfiltrates data, sends unauthorized communications, or takes unintended actions is an incident – and the containment, forensic, and notification steps are meaningfully different from those for a compromised human account.
- [ ] **Agent activity logs are accessible and structured well enough to support incident investigation without manual assembly across systems** — MED
  Incident response speed depends on log quality and accessibility. If reconstructing what an agent did requires pulling from multiple systems or escalating to a vendor, mean time to respond will be measured in hours, with direct implications for breach notification timelines.

08 Ownership

- [ ] **A specific team owns the agent's access scope and behavioral policy** — MED
  Without a designated owner, there is no accountable decision-maker for access scope changes or policy decisions. That absence is most visible during incidents and access reviews.
- [ ] **Access scope and behavioral policy are reviewed on a defined, recurring schedule** — MED
  An agent's footprint typically expands over time as new tools get connected and use cases grow. A review cadence is what catches access drift before it becomes a control gap.
- [ ] **Employees have been briefed on what the agent can and cannot do, and who to contact if something behaves unexpectedly** — LOW
  Security controls only work if users understand the basics of what they're working with. Employees need to know an agent's scope, what actions it can take on their behalf, and who to contact if something looks wrong.

09 Prompt Integrity

- [ ] **Exposure to external and user-supplied content has been assessed for prompt injection risk** — HIGH
  When an agent reads emails, documents, or web content as part of its workflow, that content can contain instructions designed to redirect its behavior. Prompt injection is one of the highest-priority attack vectors for agentic AI and is largely invisible without explicit assessment.
- [ ] **An agent's behavioral instructions are documented, approved, and have an assigned owner with a defined review process** — MED
  The system prompt governs everything an agent does in a deployment – its role, permissions, and constraints. In practice it is often written by a developer to get something working and never formally approved or assigned to an owner. Without clear ownership, there is no accountability for what the agent is instructed to do.
- [ ] **A change control process exists for reviewing and approving updates to the agent's instructions or connected tools** — LOW
  Informal changes to system prompts or tool connections accumulate without oversight. A change control process ensures updates are reviewed before taking effect, and creates a record of what changed and when.

10 Data Handling & Supply Chain

- [ ] **All MCP servers the agent connects to have been vetted — including open-source and third-party servers** — HIGH
  A compromised or malicious MCP server can exfiltrate data an agent passes through it, return manipulated results, or inject instructions into the agent's context. Each unvetted server is an uncontrolled trust boundary in the deployment.
- [ ] **Data residency and retention compliance requirements have been assessed for all data the agent processes and transmits** — MED
  An agent may process PII, financial data, or regulated content as part of its workflows. Where that data is transmitted and how long it is retained may be subject to GDPR, HIPAA, PCI DSS, applicable state privacy laws, or contractual data residency obligations.
- [ ] **High-risk actions require human approval before execution, with defined and enforced thresholds** — MED
  For actions that are irreversible or high-impact — sending external communications, modifying financial records, deleting data — a human-in-the-loop requirement is the last line of defense. Without a defined threshold, those actions execute autonomously.

If these questions are hard to answer, agents are likely operating with more access and less visibility than your security posture requires. Aembit provides the identity, access policy, and audit infrastructure to close these gaps — so security teams can say yes to agentic AI deployments with the same controls they apply to their human workforce.

Request a demo at **aembit.io**