SALT

The CISO's Checklist for Securing AI Agents and MCP

An Explainer for Security Leaders

The rapid adoption of AI agents and protocols like MCP is fundamentally reshaping our API ecosystem and, consequently, our attack surface. Industry analysts predict the vast majority of API traffic will soon be generated by AI agents, making them the primary consumers of our data and services. This checklist is designed as a strategic tool to help you assess your organization's readiness for this new reality. It will help you identify critical gaps in visibility, governance, and threat protection, providing a clear, data-driven foundation for conversations with your board and executive team about securing our next wave of innovation.

AI Agent & MCP Security Readiness Checklist

Use this checklist to evaluate your current security posture against the unique risks posed by autonomous AI agents and the APIs they consume.

1. API Discovery & Visibility

You cannot protect what you cannot see. The speed of AI development creates massive blind spots.

- [ ] Comprehensive Inventory: Do we have a continuously updated, real-time inventory of ALL APIs, including those exposed via MCP servers and third-party AI services?
- [ ] Agent Traffic Identification: Can we reliably distinguish between API traffic generated by human users versus autonomous AI agents to apply different security policies?
- [ ] Shadow API Detection: Does our process automatically discover and flag undocumented or "shadow" APIs created by new agentic workflows?

2. Secure Development & Lifecycle

Security for AI must be built in, not bolted on. This starts with how APIs are designed and tested.

- [ ] AI-Specific Abuse Case Testing: Are we testing APIs for AI-specific abuse cases, such as how a compromised agent could exploit them for data exfiltration or to bypass business logic, before they are published to an MCP server?
- [ ] "Agentic Experience" Design: Are we designing APIs specifically for an "agentic experience," with clear, atomic functions that limit the potential blast radius if an AI agent is compromised?
- [ ] MCP Exposure Governance: Do we have a formal review process to govern which APIs are exposed to AI agents and MCP servers, ensuring they are well-documented and fit for autonomous consumption?

3. Governance & Posture Management

Poorly managed access controls are a primary vector for AI agent attacks. Governance is critical.

- [ ] "Agentic Experience" Policies: Have we shifted from a "developer experience" to an "agentic experience" model, with specific, strict access control policies for machine identities?
- [ ] Least Privilege Enforcement: Can we analyze agent behavior to identify and remediate overly permissive or exposed APIs before they are exploited?
- [ ] Data Flow Governance: Do we have visibility into the specific types of data (e.g., PII, financial) that AI agents are accessing and can we enforce policies to prevent unauthorized data exposure?

4. Runtime Security & Threat Protection

Most AI-related vulnerabilities are exposed through the APIs they use. Runtime protection is essential.

- [ ] Behavioral Anomaly Detection: Do we have a baseline of normal AI agent behavior, and can we detect and alert on deviations that indicate a compromise or misuse?
- [ ] Business Logic Abuse Detection: Can our tooling detect a compromised agent performing malicious actions (like data exfiltration) that abuse the intended functionality of an API, bypassing traditional WAFs?
- [ ] Real-Time Blocking: Can we automatically block or terminate a malicious or compromised AI agent's session in real time to prevent a breach?

5. Incident Response & Forensics

When an agent is compromised, a rapid, context-rich response is crucial to limit the blast radius.

- [ ] AI-Specific Playbooks: Do our incident response playbooks include specific scenarios for a compromised AI agent or a malicious MCP server?
- [ ] Attack Timelining: In the event of an incident, can we quickly reconstruct the full timeline of an attacker's actions, from initial compromise to final objective, across multiple API calls?
- [ ] Tabletop Exercises: Do we regularly conduct tabletop exercises that simulate an AI-driven API attack to test our team's readiness?

Your Readiness Score

12-15 Checks: Strong Posture. Your program is well-prepared to govern and secure AI agents across the full lifecycle.

7-11 Checks: Partial Coverage. You have foundational pieces but significant proactive and preventative gaps remain.

0-6 Checks: High Risk. Your organization is highly exposed. Immediate action is required to address critical blind spots.