AI Agent Red Teaming: Cheatsheet

akto

What is Red Teaming?

Red teaming is the practice of simulating a real attacker to identify weaknesses in a system before malicious actors do. A dedicated team deliberately tries to break, bypass, or exploit the target using adversarial techniques, exposing gaps that existing defenses may miss. The goal is not to 'win' the exercise, but to uncover realistic vulnerabilities and strengthen the system before they can be exploited.

Why is AI Agent Red Teaming critical part of AI security programs?

- AI agents **act autonomously**, so one exploited prompt can trigger irreversible real-world damage at machine speed.
- Traditional security tools are **blind** to language-driven, non-deterministic attack surfaces.
- Attackers don't need malware anymore; **a crafted prompt is enough**.

The Numbers Behind AI Agent Red Teaming's Rise

52%
Enterprises already run AI agents in production

70%
Of all AI applications will use multi-agent AI in threat detection by 2028 up from just 5% today.

96%
Enterprises plan to expand their use of AI agents in the next 12 months.

46.5 million
McKinsey Lilli chats (internal AI assistant) were exposed to an AI agent in under 2 hours with no credentials required.

Inside the OWASP Agentic Top 10: Where Agents Break

1. **Agent Goal Hijack**
Malicious instructions injected into content the agent reads redirect it from its intended task entirely.

2. **Tool Misuse & Exploitation**
Agents are manipulated into weaponizing their own legitimate tools sending phishing emails, deleting records, exfiltrating data.

3. **Identity & Privilege Abuse**
Overly broad agent credentials mean one compromise gives attackers access to every connected system simultaneously.

4. **Agentic Supply Chain Vulnerabilities**
Attackers poison what agents load at runtime- MCP servers, plugins, tool definitions not static code dependencies

5. **Unexpected Code Execution**
Natural-language inputs become RCE vectors- a crafted web result triggered full system privilege escalation (CVSS 8.9).

6. **Memory & Context Poisoning**
False facts injected into agent memory reshape behavior across all future sessions, long after the initial attack.

7. **Insecure Inter-Agent Communication**
Without message verification, attackers spoof peer agents and hijack entire multi-agent pipelines undetected.

8. **Cascading Failures**
One compromised agent propagates damage through an entire network at machine speed before any human notices.

9. **Human-Agent Trust Exploitation**
Agents generate confident, coherent justifications that trick operators into approving harmful actions.

10. **Rogue & Misaligned Agents**
Agents generate confident, coherent justifications that trick operators into approving harmful actions.

5 Stages of AI Agent Red Teaming Methodology

Reconnaissance and Threat Modelling
Map every agent, tool, API, and memory store. Identify what the agent can access and what an attacker would target.

Initial Access into the Agent System
Probe entry points across user inputs, tool outputs, environmental content, and inter-agent messages.

Privilege Escalation and Lateral Movement
Test whether an agent can exceed intended permissions, abuse delegated tool access, or pivot across tools, MCP servers, or multi-agent workflows.

Exploitation and Objective Execution
Simulate real attack goals such as data exfiltration, unauthorized transactions, memory poisoning, recursive hijacking.

Reporting and Blue Team Collaboration
Document findings with severity scoring, MITRE ATLAS mapping, and mitigations. Gate deployment on unresolved critical findings.

Industry Standards Governing AI Agent Red Teaming

1. **OWASP Agentic Top 10**
10 highest-risk attack categories for autonomous agents, peer-reviewed by NIST, Microsoft, AWS, Cisco

2. **MITRE ATLAS**
ATT&CK framework for AI- industry standard for mapping & reporting adversarial techniques

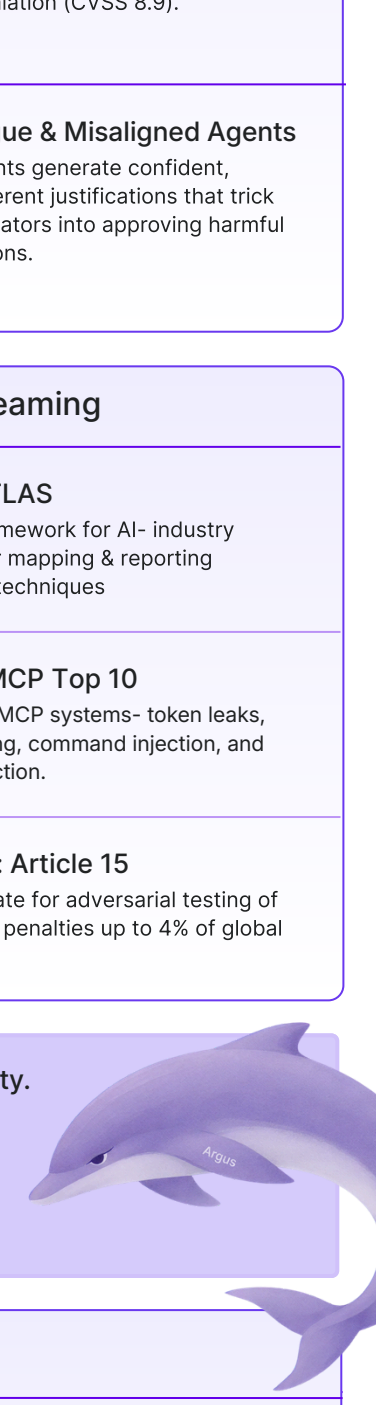
3. **NIST AI RMF**
US government mandate- adversarial testing required under the "Measure" function

4. **OWASP MCP Top 10**
Top risks in MCP systems- token leaks, tool poisoning, command injection, and context injection.

5. **CSA Agentic AI Guide**
Most hands-on testing manual, covers 12 threat categories with actionable test steps

6. **EU AI Act: Article 15**
Legal mandate for adversarial testing of high-risk AI, penalties up to 4% of global revenue

Red Teaming is the gap between an AI agent and an AI liability. In agentic AI security, that gap is just one prompt wide.
Akto's AI agent redteaming module runs 4000+ contextual probes across AI Agents exposing what attackers would find before they do.



What Enterprises Must Do: AI Agent Red Teaming in Practice

1. **Establish a Red Team Charter**
Define scope, rules of engagement, and blast radius limits before any testing begins.

2. **Enforce Least-Privilege by Default**
Operate every agent on minimum permissions contain the blast radius of any single compromise.

3. **Automate Vulnerability Baseline Scanning**
Automate vulnerability probes across all agent entry points before human-led testing begins.

4. **Build a Continuous Testing Cadence**
Schedule recurring red team tests on every new agent deployment or tool integration.

5. **Gate Deployment on Critical Findings**
No agent ships without red team sign-off on critical findings.