

MCP Security

Connecting AI agents to your tools through the Model Context Protocol opens a new attack surface. Here is what goes wrong and the one habit that keeps it safe.

WHAT IT IS

MCP is the open standard that lets AI assistants connect to your tools and data. Its weakness is that it treats a tool's plain-text description as trusted instructions, so a poisoned tool, a malicious server or a stolen token can make an agent leak data or run code on your systems.

WHY IT MATTERS

10,000+

active public MCP servers by the time Anthropic donated the protocol

Anthropic, December 2025

72.8%

tool poisoning attack success rate across 20 production MCP agents

MCPTox benchmark, 2025

CVSS 9.6

critical remote code execution in the mcp-remote proxy (CVE-2025-6514)

NVD, July 2025

WHAT HAPPENED

In September 2025 a malicious npm package called postmark-mcp copied a legitimate project, shipped 15 clean versions to earn trust, then added one line that secretly copied every outgoing email to an attacker. It was described as the first malicious MCP server found in the wild. Source: Koi Security, September 2025.

RED FLAGS TO WATCH FOR

Treat these as warning signs on any MCP server.

- A tool description that contains instructions or asks to read files
- New tools or changed descriptions after you approved a server
- A connector requesting far broader access than its job needs
- Outbound traffic to unfamiliar domains, especially after an update
- API keys stored in plaintext in an MCP config file

THE ONE RULE THAT STOPS IT

Treat every MCP server as untrusted third-party code and keep a human in the loop for any consequential action.

THEN BUILD THE HABIT AROUND IT

- Install only vetted servers from verified publishers and pin the version you reviewed
- Enforce least privilege and scope access per tenant on the server side

- Forbid token passthrough and validate token audience
- Isolate servers, log every tool call and monitor outbound traffic

MYTHS AND FACTS

MYTH

MCP is secure by default.

FACT

It standardises connection, not protection. The protocol does not stop tool poisoning or a flawed server implementation.

MYTH

If a server was safe when we approved it, it stays safe.

FACT

A rug pull can change a tool after approval, and name-based trust will not re-check it.

MYTH

A popular open-source server is safe to install.

FACT

postmark-mcp had 15 clean releases before one poisoned line, so popularity is not provenance.

THE LEGAL DUTY

Under NIS2 and Sweden's Cybersäkerhetslagen, MCP servers are third-party components covered by supply-chain security and incident-reporting duties, and a cross-tenant leak can trigger GDPR notification to IMY within 72 hours.

GET HELP

Connecting AI agents to your systems? Talk to eBuilder Security about governing MCP safely.

[Read the Full Article →](#)