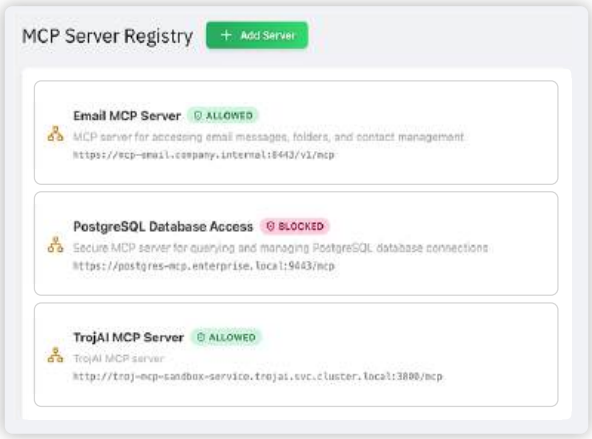
TROJ.AI

TrojAI Defend

for MCP

SOLUTION BRIEF



MCP runtime security for the enterprise

Model Context Protocol (MCP) is an open protocol that allows AI agents to connect with external data, tools, and services in a standardized way. While MCP gives AI agents the power to act autonomously, it also raises a new class of risk. Each MCP server, tool, and connection introduces another moving part that traditional security tools weren't built to monitor or manage, opening the door to new attack vectors.

TrojAI Defend for MCP delivers complete visibility and control over MCP workflows. It provides a registry for MCP servers and tools, unmasks malicious tool descriptions that could contain prompt injections or other attacks, and tracks tool integrity over time. TrojAI Defend for MCP enables enterprises to innovate with agentic AI securely.

Discover and secure every MCP server

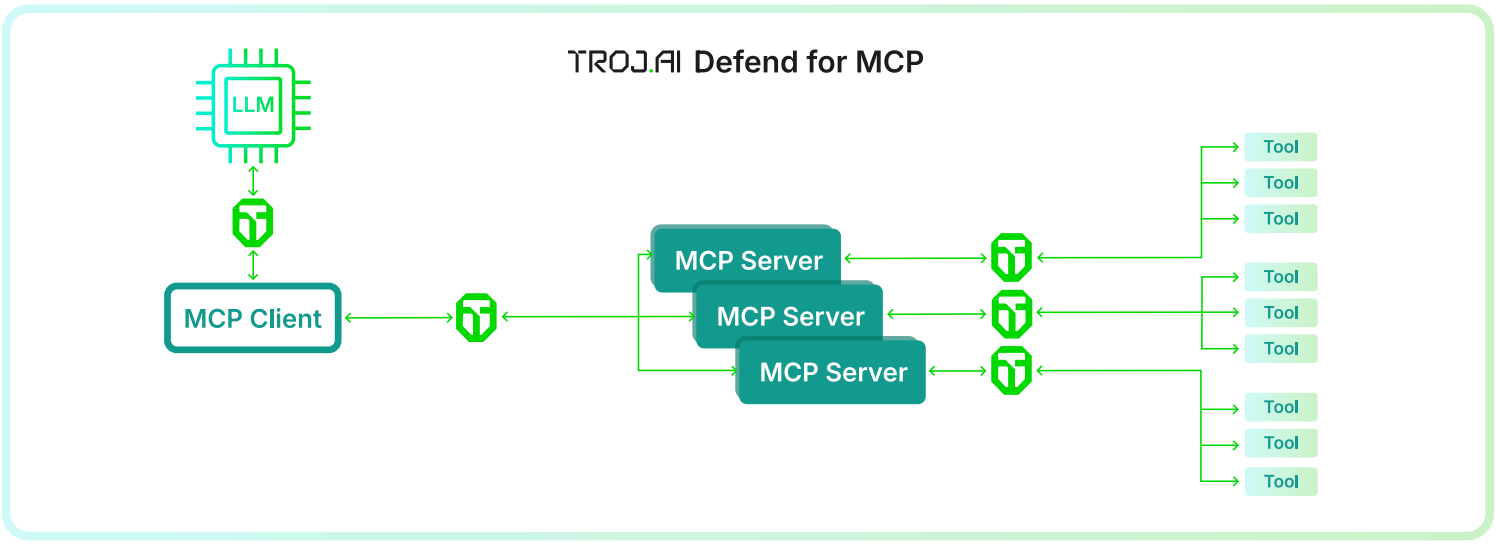
As organizations rapidly adopt MCP to power agentic AI workflows, knowing which servers are active and whether they're safe has become a monumental challenge. These risks are compounded given that each MCP server can call multiple unvetted tools.

TrojAI Defend for MCP discovers and registers all MCP servers to stop unauthorized access by malicious actors and eliminate shadow MCP. MCP servers and their related tools can then be approved in the registry, preventing rogue or unauthorized use of risky tools.

Stop rogue MCP servers and prevent manipulation attacks

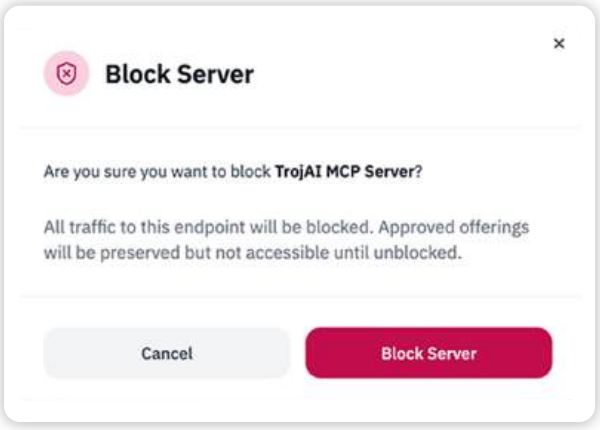
Malicious or unverified servers can expose tools that perform unauthorized actions or leak sensitive data. Attackers can hide prompt injections inside tool metadata — names, descriptions, or parameters that seem harmless but change how an AI behaves.

TrojAI Defend for MCP identifies all MCP traffic to and from each server to block unregistered or rogue servers. Hidden communication paths are eliminated to protect against common AI attacks like prompt injection, data exfiltration, and more.



Instantly detect tampering

Even approved MCP servers and tools can be tampered with, creating hidden risks within trusted workflows. TrojAI Defend for MCP continuously verifies every server and tool definition, detecting any drift or modification in real time. When unauthorized changes appear, TrojAI Defend for MCP isolates and blocks the tool, stopping tampering, poisoning, or rug-pull attacks before they spread.



AI policies built to defend MCP workflows

As MCP becomes the backbone of agentic AI, enterprises need policies that understand how agents actually behave. Without MCP-specific guardrails, prompt injection, data leakage, and unauthorized tool use can slip through unseen.

TrojAI Defend for MCP enforces comprehensive, customizable policies that inspect, audit, and secure every agent interaction. These policies strengthen governance by ensuring all activity adheres to enterprise data handling rules, and they maintain detailed audit trails that provide the evidence needed for compliance and incident response.

About TrojAI

TrojAI is a comprehensive AI security platform that protects AI applications, models, and agents. The best-in-class platform empowers enterprises to safeguard AI systems both at build time and run time. TrojAI Detect automatically red teams AI models, safeguarding model behavior and delivering remediation guidance prior to deployment. TrojAI Defend is an AI application firewall that protects enterprises from real-time threats. Built by data scientists and cybersecurity experts, TrojAI secures the largest enterprises with a highly scalable, performant, and extensible solution.

Learn more at Troj.ai