Prisma AIRS AI Agent Security

A Unified Foundation for Securing Agents

Imagine hiring hundreds of new employees, each with access to your tools, data, and systems—with no clear way to track what they do or to stop them if they go rogue. That's the reality enterprises face as AI agents and the Model Context Protocol (MCP) servers that power them spread across SaaS apps, low-code/no-code platforms, and custom builds.

The result is two core challenges. The first one is agent monitoring in which agents are deployed everywhere across the organization without centralized visibility or control. The second challenge is lifecycle risk, where each agent introduces overpermissive identities and unpredictable runtime behaviors that traditional security tools can't contain.

Key Benefits

- **Universal visibility:** Gain centralized visibility into all AI agents, regardless of whether they are SaaS, custom built, or third party.
- **Enterprise-wide enforcement:** Unify your security posture by enforcing consistent policies from a single, centralized platform.
- **End-to-end protection:** Protect against the full range of AI-specific threats throughout the agent lifecycle, from initial configuration to runtime operation.
- **AI-specific threat prevention:** Actively prevent memory manipulation, tool misuse, and other adversarial attacks that target an agent's logic.
- **Granular data control:** Ensure compliance and privacy by providing fine-grained control over what data an agent can access, use, and share.

Prisma® AIRS™ AI Agent Security addresses both challenges:

- Secures every AI agent across its lifecycle, from configuration to runtime.
- Monitors AI agents across the organization with centralized visibility, policy, and control.

We offer a single platform for visibility, control, and consistent policy enforcement. With built-in defenses against AI-specific threats, Prisma AIRS agentic security focuses on prompt injection and data leakage, as well as provides fine-grained access and full audit trails.

Integrated Security Components

Prisma AIRS AI Agent Security compromises three core systems.

Table 1. Core Systems

Component	Purpose
SaaS AI Agents	Secures AI agents in third-party SaaS platforms by providing full lifecycle protection—from configuration and permissions to real-time monitoring and threat prevention.
MCP Relay	Monitors agents with a centralized proxy that enforces consistent security and visibility across every MCP workflow.
Prisma AIRS Managed MCP Server	Gives developers a lightweight, on-demand runtime protection that integrates seamlessly into any agent workflow without slowing innovation.

Capabilities

Secure SaaS Agents End to End

Enterprises are rapidly adopting SaaS AI agents, such as Microsoft Copilot, because they deliver immediate productivity gains without needing to build or manage custom AI. They come preintegrated with critical business apps, like Microsoft Office 365, Teams, and Dynamics, making them attractive for fast adoption.

But these agents run as black boxes inside SaaS platforms. Security teams cannot see how inputs are processed, what intermediate steps are taken, or which external APIs are invoked. Traditional security tools lack visibility at this layer, leaving organizations blind to runtime behavior. That means an overprivileged agent could silently access sensitive SharePoint files, leak Teams conversations, or be manipulated through prompt injection.

Our SaaS AI Agent protection closes that gap. It's designed to enforce least-privileged access, continuously monitor activity, and block risky permissions, privilege escalations, or data leaks in real time. Every action is logged for compliance, traceability, and forensic analysis—bringing visibility and control back to an otherwise opaque SaaS execution layer.

This ensures that Copilot Studio and similar SaaS agents deliver productivity gains without introducing uncontrolled security and compliance risks.

Manage Agents with a Mandatory Gateway

MCP-based systems bring enormous flexibility to AI agent development, but their decentralized nature also introduces security vulnerabilities. Each agent interacts dynamically with multiple MCP servers, tools, and data sources—some enterprise owned and some third party—creating new vectors for threats, such as prompt injection, malicious URLs, context manipulation, and credential leakage.

MCP Relay

The MCP relay acts as a proxy between MCP clients and servers to enforce runtime security for every interaction. Each time an agent communicates with a server or tool, the relay calls the Prisma AIRS Runtime API to apply advanced security checks. This ensures that malicious instructions, poisoned tool descriptions, or exposed credentials are detected and blocked before they can affect workflows.

These protections include detection of both context poisoning through manipulated tool metadata and credential leakage through exposed parameters during tool calls. The relay enforces these runtime inspections consistently across all agentic workflows so every tool invocation is validated and secure.

Managed MCP Server

The Managed MCP Server extends this protection by providing a hosted runtime enforcement layer. It works seamlessly with any MCP-compatible agent, regardless of model or platform. Plus, it also leverages all Prisma AIRS detection capabilities—including protection against over 31 prompt injection techniques, topic abuse, harmful content, and web- or DNS-based attacks.

MCP Relay and Managed MCP Server Together

When paired, the MCP Relay and Managed MCP Server provide end-to-end runtime protection. The relay enforces security at every interaction point, and the managed server delivers AI runtime defense on demand. Together, they give enterprises a unified, fully managed security foundation for MCP-based AI ecosystems—ensuring every agent runs safely and predictably across diverse environments.

Agent Protection on Demand

While enterprises require centralized governance at scale, developers need security that matches the speed of innovation. AI agents are built across diverse platforms, and traditional security tools are too slow, rigid, or infrastructure-heavy to keep pace.

The Managed MCP Server delivers on-demand, developer-friendly protection. You can invoke it directly inside workflows without rewriting code or standing up infrastructure. It acts as a universal checkpoint for any MCP-compatible agent.

When called, the server inspects each tool invocation, prompt, and model response in real time to detect and block critical threats:

- **Prompt injection:** Prevents malicious instructions from altering agent behavior.
- **Context poisoning:** Stops manipulation of tool descriptions and ensures reliable inputs.
- **Malicious URLs and data leaks:** Blocks risky links and prevents sensitive data exfiltration.
- **Privilege escalation:** Protects against unauthorized access to underlying systems.

This "security-on-demand" model lets developers secure both experiments and production agents with minimal friction. By embedding runtime validation into development workflows, it extends consistent enterprise-grade protection to every agent from the moment it is created—without slowing innovation.

Building and Consuming Agents—Secured in One Platform

Enterprises adopt AI agents in two ways: by consuming prebuilt agents inside SaaS platforms like Salesforce or Microsoft, and by building custom agents in-house to integrate proprietary data and workflows.

Most organizations do both—using SaaS copilots for speed and reach, while developing custom agents for differentiation and control. Prisma AIRS secures both models in one platform, giving you flexibility to innovate without creating new blind spots. The result is a unified security posture across your entire AI ecosystem, providing a consistent layer of protection regardless of where or how your agents were deployed.

For Agents You Consume

SaaS AI Agent Security enforces least-privileged access, data boundaries, and audit controls so embedded copilots in platforms, like Salesforce or Microsoft, operate safely. At runtime, the MCP Relay inspects API and tool calls, blocking privilege escalation and data exfiltration in real time.

For Agents You Build

For in-house development, the Managed MCP Server integrates directly into developer workflows, allowing you "shift-left" security. By providing on-demand validation of code and tool calls, it ensures that security is a core part of the build process. The MCP Relay, meanwhile, acts as the universal security checkpoint for all agent traffic, securing the agent's communications to all destinations.

Portability and Deployment Flexibility

By decoupling security from the underlying infrastructure, we deliver a portable and scalable solution that provides comprehensive protection across the decentralized AI agent landscape.

- **SaaS and third-party agents:** Secure agents are embedded within third-party SaaS applications, where traditional security tools lack visibility. Our solution integrates seamlessly to monitor and protect agent activities without requiring direct access to the underlying cloud infrastructure.

  Secure AI agents built on modern low-code/no-code platforms (e.g., Microsoft Power Apps and Google AppSheet) often bypass traditional security controls. Our solution extends security policies to these platforms, ensuring that even nontechnical users cannot create or deploy insecure agents.
- **Custom-built and internal agents:** Secure agents and models are developed in-house, whether they are running on-premises, in a private cloud, or within a public cloud environment (Amazon Web Services, Azure, or Google Cloud). Our platform provides consistent security and governance for all custom-built agents, ensuring they adhere to enterprise-wide security standards from the moment of their creation.

This universal deployment capability provides the horizontal coverage necessary to manage the proliferation of AI agents across your organization, from structured IT environments to decentralized, user-driven deployments. It allows you to protect your business without dictating your technology stack.

Flexible Consumption Model

Customers can acquire or use their existing Flexible Software NGFW (FW-Flex) credits to deploy Prisma AIRS AI Runtime Security instances from Strata™ Cloud Manager and reconfigure security quickly with minimal procurement roadblocks. For API-based instances, developers can create a deployment profile for their FW-Flex credits, which allows them to create an API key in Strata Cloud Manager. They can then use this key to make API calls from application code.

Product Specifications

For cloud-specific AI model support on network-based intercepts, check the AI Models Support table. For capacity and throughput information, see the VM-Series Virtual Next-Generation Firewalls datasheet.

Additional Resources

- Prisma AIRS product page
- Software NGFW Credit Estimator
- Software Firewall Selector

paloalto NETWORKS

3000 Tannery Way
Santa Clara, CA 95054

Main: +1.408.753.4000
Sales: +1.866.320.4788
Support: +1.866.898.9087

www.paloaltonetworks.com

prisma_ds_prisma-airs-ai-agent-security_102825