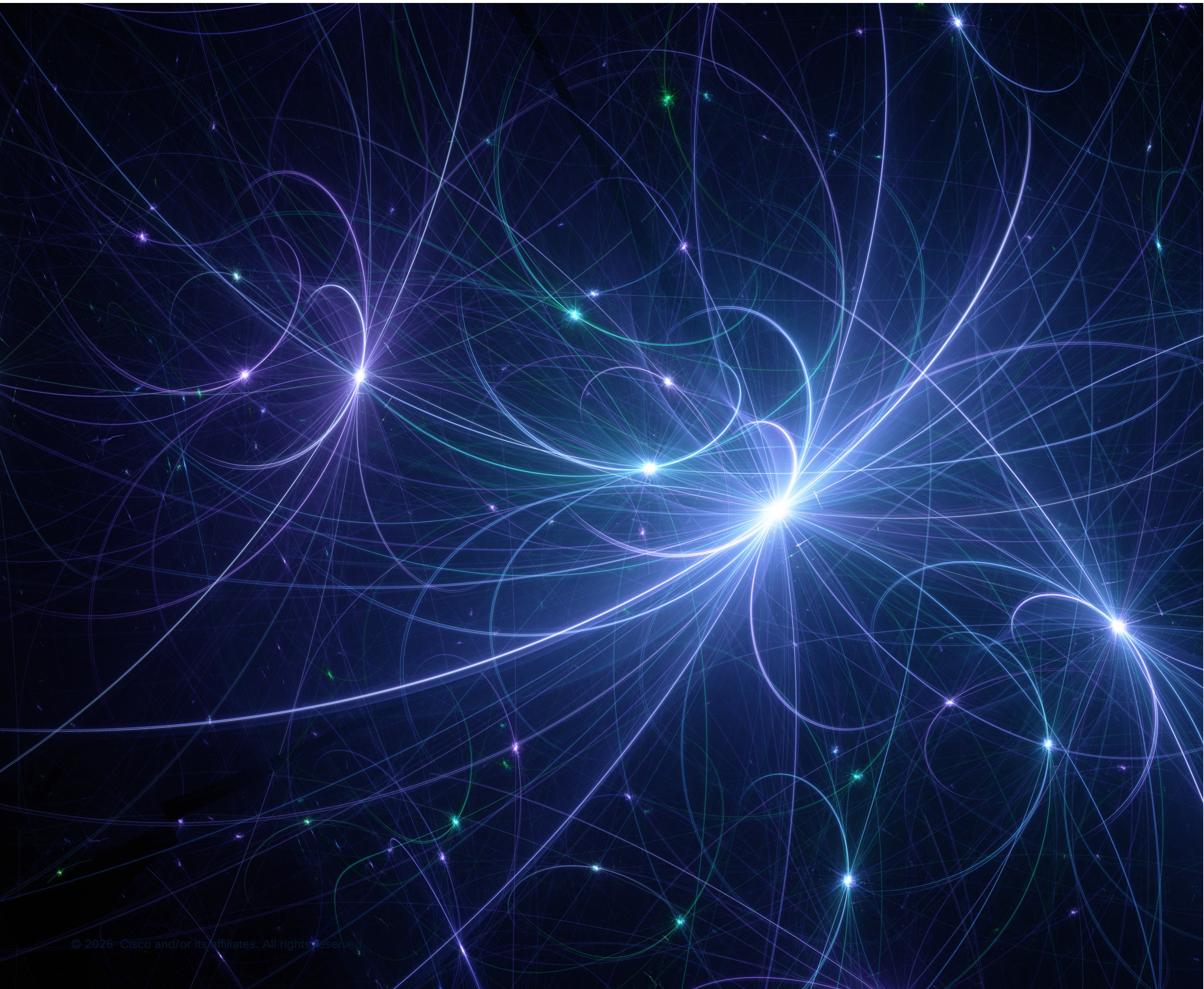
Solution Brief Cisco Public

Zero Trust for Your Agentic AI Workforce



Agentic AI represents a new digital workforce that autonomously performs tasks, makes decisions, and operates at machine speed and scale, but unlike humans lack common sense and often disregard consequences. While enterprises rapidly adaopt these AI agents, securing them poses unique challenges.

Cisco Zero Trust for Agentic AI focuses on protecting the world from AI agents by extending Zero Trust principles beyond humans to govern AI agents' identity, access, and behavior. This ensures AI agents work securely across enterprise environments without becoming security liabilities.

Challenges

As AI is in its nascent stages with constant evolution of the technology, it is a wild wild west scenario in the AI ecosystem. Agents are being deployed everywhere by everyone. At the same time, as agentic AI systems act autonomously, they connect to APIs, invoke tools, access enterprise data, and coordinate actions across multiple enterprise systems and services. This is underscored by challenges unique to AI agents and the tools and resources these agents access:

- **Agents are everywhere:** Today, different types of agents are being deployed by enterprises by different types of users. This creates a challenge into the visibility of agents interacting with the environments and their activities resulting in shadow agents doing any activity as they please
- **Fragmented Enforcement:** Agents are being deployed across different planes resulting in policy enforcement being spread across the planes. On top, these enforcement points are controlled by IT teams, Dev teams and hardly ever by security teams. This creates security blind spots that agents can easily take advantage of whenever they please.
- **Unpredictability of agent actions:** AI agents act unpredictably, lacking human judgement, defying static security policies, and complicating behavior control creating havoc in enterprise environments.

Existing tools lack integrated identity and access controls for AI agents, creating gaps in visibility and enforcement. What happens when you authenticate an agent, give it access to tools with sensitive information, and then it goes rogue?

Solution: Zero Trust Access for Agentic AI

Organizations need a new approach built for visibility, protection, and control as AI systems become embedded across enterprise operations. To address these challenges, Cisco applies Zero Trust principles across identity, access, and AI behavior, creating a unified security framework for agentic AI.

- **Know every agent** – Establish trust with AI agents.
- **Authorize every action** – Control who/what/when/how of systems, tools, and data access.
- **Adapt to risk in real time** – Monitor and enforce safe AI actions at runtime.

Know every agent

Identity is the foundation of Zero Trust. Organizations must verify every entity interacting with AI systems, including human users, devices, and non-human identities such as AI agents. Without strong identity assurance, organizations cannot reliably determine who or what is initiating actions across AI workflows.

Cisco Duo allows us to know every agent by providing identity assurance through agent visibility & identity management. Discover, register, and manage AI agents, MCP servers, and tools, mapping agents to accountable humans for auditability and assigning access policies to agents.

Feature	Description
Agent discovery	Discover all agents, MCP servers, and tools in use across enterprise environment.
Agent directory	Register agents into a directory with dynamic inventory checks to identity unauthorized entities. Map agents back to accountable humans for auditability.
Access policies for agents	Assign access policies on which agents can access what tools, resources and data
Agent lifecycle management	Discover all agents, MCP servers, and tools in use across enterprise environment.

Authorize every action

Without strong access governance, AI agents and applications can use excessive permissions, invoke unintended services, or interact with sensitive data or resources beyond their intended scope. If an agent is compromised or just tries to go beyond its original purpose due to its non-deterministic nature, broad access to sensitive resources can lead to major problems. Organizations need to defend against agent behavior drift or unintentional actions when AI agents access resources.

Cisco Secure Access allows you to authorize every agent action with fine-grained access control by enforcing identity-aware, intent-based policies on AI agents' access to MCP servers, tools, and resources. By applying least-privilege principles and intent-based access controls, Secure Access ensures that AI systems interact only with approved tools, data sources, and services for the right reason and length of time.

Feature	Description
Fine-grained authorization policies	Enforce precise, least-privilege policies for agent tool use and actions based on agent identity and context.
Time-bound access controls	Designate short-lived tokens to agents to restrict the time an agent has access to tools and resources.
MCP gateway for policy enforcement	Direct traffic through an MCP gateway for consistent, simplified enforcement.

Adapt to risk in real time

Even trusted identities operating with approved access can behave in unexpected ways. Agentic AI systems make decisions, invoke tools, and interact with enterprise systems autonomously, which means their actions must be continuously monitored and governed.

Runtime oversight is essential to ensure AI systems operate safely, comply with policy, and do not expose sensitive data or trigger unintended actions.

Cisco AI Defense provides the ability to adapt to risk with real-time behavior monitoring and protection for AI interactions, applying safety and security guardrails that defends against threats like prompt injections, unsafe outputs, and anomalous agent behavior before they can impact systems or data.

Feature	Description
Security guardrails	Secure interactions between agents, MCP servers, and tools, and defend against threats like prompt injections.
Safety guardrails	Guard against sensitive data leaks or inappropriate content in agentic interactions.
Intent-based monitoring and control	Understand intent and context of agentic interactions to ensure agents are operating in a safe, intended manner.
Web threat protection	Block agent access to malicious websites.

A Unified AI Security Foundation

Security for AI demands more than point solutions. It requires an integrated approach that continuously checks identity context along with intended access and safe behavior for end-to-end protection. Cisco delivers an integrated architecture so organizations can confidently adopt agentic AI.

Cisco Zero Trust for Agentic AI empowers organizations to securely harness the power of AI agents while maintaining strict governance, accountability, and protection against emerging threats.

To learn more, visit: https://cisco.com/go/securing-agentic-ai.