Zero Trust for AI Agents: The Only Model Built for What's Next

June 16, 2026

Introduction

For the last two years, most enterprise AI conversations have focused on productivity. Faster research. More accurate writing. Better assistance for employees. That was the opening chapter.

What comes next is more consequential: AI that does not just generate output, but takes action.

AI agents are beginning to access applications, use tools, retrieve data, trigger workflows, and act on behalf of users. As Anthropic notes in its Zero Trust for AI Agents white paper, agentic systems introduce a different trust surface because they can "interpret goals, select tools, and execute multi-step operations." [1] That shift matters. Once AI moves from answering questions to operating inside the business, the challenge is no longer just how to use AI productively. It is how to govern systems that can act with speed, autonomy, and reach.

That is why this moment does not call for a new security theory. It calls for applying the right one with more discipline. Zero Trust was built for environments where trust cannot be assumed. That is exactly why it is the ultimate answer to securing agentic-powered enterprises of the future.

As AI agents gain autonomy, enterprises need a proven security model for governing access, action, and trust.

How do AI agents change the nature of risk?

An AI assistant that summarizes a document creates one kind of risk. An AI agent that can query a database, update a ticket, call an API, move data between systems, or trigger a downstream workflow creates another.

The difference is agency.

When AI moves from generating content to taking action, security teams have to think beyond model outputs and prompt controls. They have to think about operational behavior: what the agent can access, what it can do with that access, what tools it can invoke, what systems it can interact with, and how those actions are governed in real time. That is why agent security is not just an AI discussion. It is an architecture discussion.

Any entity that can access business systems, interact with sensitive data, or take action across workflows must be governed accordingly. It needs a verified identity. It needs tightly scoped permissions. Its actions need to be constrained. Its behavior needs to be visible and traceable. And its communications with applications, data, tools, and other agents need to be controlled in real time.

This is not a side issue within AI. It is a trust, access, and control problem at enterprise scale.

How do you adopt AI without letting risk outpace governance?

This is where customer conversations are heading now:

- How does Zero Trust apply to AI agents?
- Are AI agents just another application risk, or something fundamentally different?
- What changes when AI can take action instead of just generating content?
- How should enterprises think about access for nonhuman actors?
- How do we enable AI innovation without creating uncontrolled risk?

These are the right questions because agentic AI changes the operating environment.

Agents may use valid credentials. They may interact with approved systems. They may appear to be carrying out legitimate business functions. Yet they can still create risk if they are over-permissioned, loosely governed, or allowed to operate too broadly across the environment with too little visibility. That is where older trust models start to break down.

If the architecture still assumes that being on the network, inside the environment, or behind a security boundary is enough to justify access, then AI agents are not just another use case. They are a stress test for the limits of implicit trust.

Zero Trust starts with the right premise

Zero Trust is not a feature or a repackaged legacy control. It is a battle-tested security model built for environments where trust must be continuously earned and verified, which is exactly why it fits in the age of AI agents. An agent may have a valid identity, act on a user's behalf, and use approved tools, but that still should not translate into broad or persistent trust. Every connection must be explicitly verified, every access decision evaluated in context, every privilege tightly scoped, and every action visible and governable. That is not a new doctrine; it is the proven model for governing what comes next.

In the age of AI agents, access control must evolve into action control

The key question is no longer just what an identity can access, but what an agent is allowed to do once access is granted. That means defining and enforcing guardrails around:

- Which tools an agent can invoke
- Which tasks it can perform
- The condition under which it can act how often it can operate
- Whether it can delegate
- When human approval is required

Identity still matters, but identity alone is not enough. Enterprises also need runtime governance, behavioral guardrails, and full traceability. If an organization cannot tie an agent's action back to the policy, context, tool, and authority that permitted them, it does not have meaningful control.

What comes next demands stronger controls, not softer ones

In an AI-driven environment, controls that merely add friction without materially reducing exposure are not enough. Machine-speed actors are far less constrained by inconvenience than humans are, which makes architectural security more important than ever. The stronger model is built on verified identity, short-lived credentials, tightly scoped access, controlled tool use, continuous inspection, and architectures that reduce exposure in the first place. That is the core of secure AI agent adoption:

Principle	Why it matters
Verifiable identity for every agent	If an agent can act inside the business, it cannot operate as an anonymous or loosely governed process.
Specific, least-privileged access	Agents should get access only to the apps, data, and workflows required for a defined task.
Constrained action, not open-ended autonomy	Approved access does not mean unlimited permission to act, invoke tools, or move data
Continuous visibility and traceability	Security teams need to know what the agent did, what it touched, and what policy allowed it.
Architecture that reduces exposure	The fewer reachable paths and exposed services, the less opportunity for machine-speed abuse.

The path forward

Enterprises do not need to lower their standards to move faster with AI. They need a security model that can keep pace with how the business is actually changing.

That means moving beyond perimeter-era assumptions. It means treating agent security as an architectural issue, not a feature checklist. It means reducing attack surface, eliminating unnecessary exposure, and making every access decision explicit. And it means applying Zero Trust the way it was meant to be applied: as a durable model for environments where trust must be earned continuously.

AI agents are changing how work gets done. They should also clarify what secure adoption really requires. Not a bolt-on control. Not a repackaged legacy model. But a proven architecture for governing what comes next.

The timing of Anthropic's publication is not coincidental, it is confirming. As the industry's leading voices in responsible AI development signal that Zero Trust is the right framework for the agentic era, Zscaler is proud to show what that framework looks like in practice. At ZenithLive '26 last week, we unveiled the industry's first complete Zero Trust platform for Agentic AI; not a roadmap, not a proof of concept, but a proven architecture built for this moment. What Anthropic describes as the right model, Zscaler delivers as a deployable reality. And that is exactly what Zero Trust was built to do.

Ready to see the Zero Trust platform for Agentic AI in action? Learn more and schedule a demo.