Outerlimit

Extending Zero Trust to the Agent Action Layer

Delivering zero credential exposure and multi-hop chain integrity with a single policy

Outerlimit
info@outerlimit.com

Contents

The New Zero Trust Layer

Zero trust is the dominant security discipline of the last decade. It rests on three principles: **never trust always verify, assume breach, and least privilege**. As a result, access decisions move from network position to verified identity, and the defense perimeter is replaced by per-request verification against an identity-bound policy.

Whilst zero trust extended access decisions to the identity layer, it did not extend execution decisions to the agent action layer.

The agent action layer is where an agent's authority becomes an effect on a connected system. Tool calls execute, APIs are invoked, and databases are modified. The decisions made at this layer are not access decisions, but instead execution decisions requiring architectural primitives that the identity layer was never designed to provide.

The risk at this layer is wider than compromise. An agent is non-deterministic and goal-directed: it chooses its actions to reach an objective, and those choices are not fixed in advance. It also acts at machine speed, taking and chaining actions faster than a person can review them. Harmful action therefore does not require an attacker. A verified agent operating on legitimate access can pursue an objective through actions no one intended, whether steered there by prompt injection, carried there by goal drift, or arriving there through its own reasoning. Compromise is one path to a harmful action, not the boundary of the problem.

This is what agentic AI changes for Zero Trust, whose three principles each assume a principal an agent is not.

Never trust always verify assumes a principal whose identity can be verified per request. An agent's identity can be verified but its intent cannot. A verified agent under prompt injection or goal drift issues a request that passes every identity check while serving an objective the enterprise never authorized.

Assume breach designs for compromised infrastructure, networks, and credentials, which are all external to the trusted principal. With an agent, the trusted principal is itself the breach vector when it takes the unauthorized action through fully authorized channels.

Least privilege assumes that scoping a principal's access also bounds its behavior. For a human or a service, it largely does, whereas for an agent it does not, because within the same granted scope a non-deterministic agent can take a far wider range of actions than the privilege grant anticipated. OWASP gives this its own name: Least Agency, the mitigation framing for Excessive Agency (LLM06). Where Least Privilege bounds what an identity may access, Least Agency bounds what an agent is permitted to do, because a verified agent acting within its granted scope can still pursue an objective no one authorized. Bounding agency, unlike scoping access, can only be done where the agent acts.

Multi-hop delegation compounds all three. When an agent calls another agent that calls a tool, the action that executes is several hops removed from the principal that initiated it, and verifying each hop in isolation does not establish that the chain is intact. An agent that has been authenticated, authorized, and granted access through the standard zero trust controls can still take consequential action at the agent action layer. The controls that govern the network and identity layers are necessary but not sufficient to extend zero trust to it. The agent action layer requires its own architectural commitment.

Anthropic's Zero Trust for AI Agents Framework[1], published in May 2026, establishes the capability requirements for zero trust at this layer. This document describes how Outerlimit's architecture produces these capabilities, enabling organizations to extend zero trust to the agent action layer.

[1]Anthropic Framework: Zero Trust for AI agents. 'A security framework for deploying autonomous AI agents in the enterprise'. https://claude.com/blog/zero-trust-for-ai-agents

Zero Trust Requirements at the Agent Action Layer

To apply traditional identity-based zero trust principles to the agent action layer, organizations must produce comparable structural properties at the point of execution.

- **Never trust, always verify**: the action must be bound to the verified identity of the calling agent at the point of execution, and in multi-agent chains, the integrity of the chain must be enforced architecturally rather than verified per hop.
- **Assume breach**: the credential that authorizes the action must not exist as a standing artifact the agent or any intermediary holds.
- **Least privilege**: the architectural baseline provides temporal and principal least privilege (single-use, identity-bound credentials), and operational policy completes the application by scoping what each agent's authority can do, how often, and where.

The decision to allow or deny must also be deterministic, and the evidence of each decision must be produced on the control path.

These properties are not produced by stronger authentication, finer-grained role assignment, or more aggressive token rotation. Those are identity-layer controls and they govern access, not execution. Outerlimit consumes the verified identity those controls establish and binds enforcement to it. It does not issue or replace identity. The agent action layer requires architectural primitives the identity layer does not provide. The properties this requires, and the architecture that produces them is described in the sections that follow.

Outerlimit Zero Credential Exposure

The standard Zero Trust model assumes credentials exist as standing artifacts somewhere. For example, an API key in an environment variable, a token in a vault or an OAuth refresh token in a session store. Every security control built around standing credentials is a control built around an asset that can be compromised. Hardening that perimeter is typically handled by credential management solutions.

In contrast, a core property of the Outerlimit architecture is zero credential exposure. This approach eliminates the perimeter rather than hardening it. The agent holds a cryptographic imprint, mathematically inert on its own. The imprint encodes the binding required for credential reconstruction, specifically the identity of the calling agent and the tool being called. The real credential is reconstructed at the enforcement point, used for the upstream call and then discarded. It does not persist anywhere as a usable secret.

As a result:

- A compromised agent yields nothing exploitable, because the agent never held the credential.
- A compromised intermediary yields nothing exploitable, because the intermediary never held the credential.
- A compromised enforcement point yields nothing exploitable across other agents, because the credential at that point was reconstructed for one call and discarded.

This is structurally different from credential management, which hardens the storage and rotation of standing secrets. Zero credential exposure removes the secret as a standing object. The class of breach that begins with credential exfiltration from an agent runtime is removed by construction, not reduced by hardening.

As Anthropic's Zero Trust for AI Agents Framework[1] outlines, the test should be "is the attack impossible, or just tedious?". Cryptographic enforcement satisfies the first.

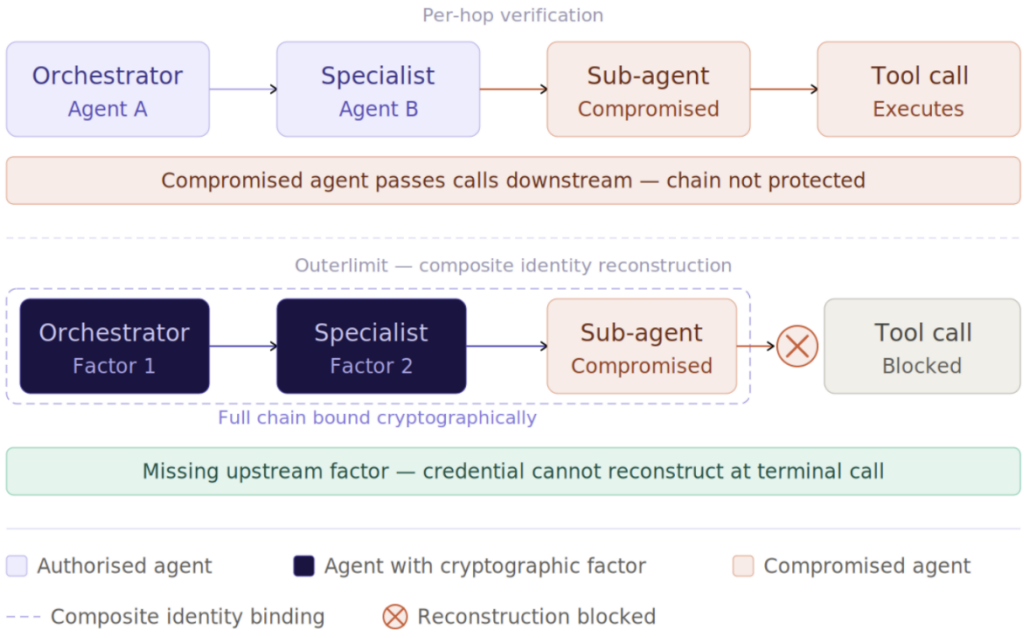
[1]Anthropic Framework: Zero Trust for AI agents. 'A security framework for deploying autonomous AI agents in the enterprise'. https://claude.com/blog/zero-trust-for-ai-agents

Outerlimit Multi-Hop Chain Integrity

In multi-agent architectures, an agent calls agent A, which calls agent B, which calls a tool. The number of hops reflects the growing sophistication of agentic deployments as orchestrator agents delegate to specialist agents, specialist agents call into sub-agents and sub-agents invoke tools. Each hop is an enforcement point in principle. The architectural question is what the enforcement decision is bound to.

Outerlimit's cryptographic enforcement binds the final credential reconstruction to the composite identity of the full chain. Every agent in the chain contributes a cryptographic factor to the composite identity used at the terminal reconstruction. If any identity in the chain is wrong, missing, or substituted, the credential does not correctly reconstruct. The check is not whether each hop verified the next but instead, whether the credential at the terminal call can be produced from the contributions of every hop that should be in the chain.

As an example, an attacker compromises an agent in a three-hop chain. Under per-hop verification, the attacker controls a verified agent and can forward calls of their choice downstream. Under Outerlimit's composite identity reconstruction, the attacker controls the intermediate agent but cannot synthesize the cryptographic contribution that the upstream agent's identity makes to the final reconstruction. The intermediate agent alone cannot produce the credential and therefore, the compromise stops at the compromised agent.

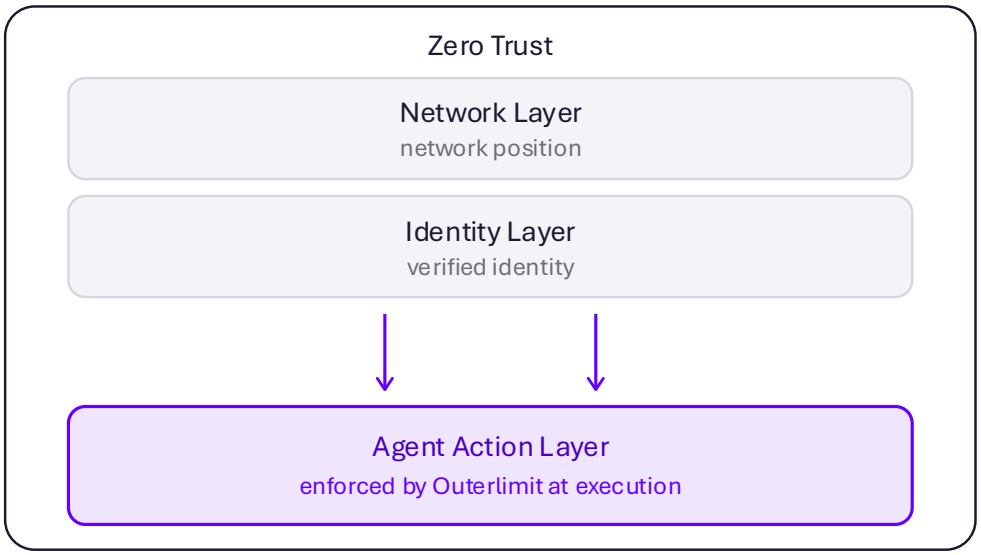


Extending Zero Trust to the Agent Action Layer

Together, zero credential exposure and multi-hop chain integrity produce the four structural properties required to extend zero trust to the agent action layer:

1. Identity-bound enforcement at the point of execution.
2. No standing-credential artifact.
3. Architecturally enforced chain integrity.
4. Deterministic decisions with evidence on the control path.

Alternative approaches that bind enforcement to observed behavior or to network path produce none of these properties, because the binding is to something other than identity.



The Outerlimit Enforcement Architecture

Zero credential exposure and multi-hop chain integrity are both consequences of a single enforcement mechanism. This section describes that mechanism and where it runs.

The decision is the reconstruction

Conventional enforcement separates the decision from the action. A policy decision point evaluates a request, returns an allow or deny and a separate enforcement point carries it out. The decision and the credential are distinct, allowing the policy check to complete before a usable credential is released to the approved caller. Anything that bypasses, misconfigures, or fails the decision step leaves the credential intact and usable.

In contrast, Outerlimit does not separate the two. The allow-or-deny decision and the production of the credential are the same event. An imprint correctly reconstructs into a usable credential only when the calling agent's identity and the requested tool match the policy bound into it. In a multi-agent chain, this occurs only when the composite identity of the full chain is present. There is no decision step that can be bypassed independently of the credential because the decision is expressed in whether the credential reconstructs correctly at all. Policy is enforced deterministically by the reconstruction, not by a guard that sits in front of it, because the decision is simply whether the credential reconstructs against identity and policy, it does not depend on why the action was attempted. A compromised agent, an injected instruction, goal drift, and the agent's own reasoning all meet the same test, and an off-policy action fails it in every case.

This is what produces the four structural properties at once: identity-bound enforcement at the moment of execution, the absence of a standing credential, architecturally enforced chain integrity, and a deterministic decision.

The point of enforcement

With Outerlimit, enforcement runs in-tenant at the agent action boundary without requiring changes to the agent or existing infrastructure.

Using this approach, there is no central decision service that requests are routed to. Instead, each enforcement point holds the policy it needs, pushed to it ahead of time and

reconstructed locally. The decision is made where the action occurs, inside the customer's environment and the request does not leave the tenant to be authorized.

Local enforcement, no central dependency

Applying zero trust at every point an agent acts is the correct goal, but the conventional approach, checking every node that calls a central authority before proceeding, has two critical trade-offs. Firstly, every action waits on a round-trip to the decision service and secondly, that service becomes a dependency that must be reachable for any action to occur. As a result, verification adds significant latency on every call and introduces a single point of failure.

Outerlimit's architecture avoids both trade-offs, because the decision is made locally at the node, as part of the action itself rather than a central authority that must be queried before each action. Since the decision is the reconstruction, there is no round-trip to a separate decision point on the path of a call. Since policy is held locally at each enforcement point, no central service must be live for an action to be authorized. Therefore, an interruption to central components does not impede enforcement. Per-node verification is achieved without the per-node coordination cost that makes it impractical elsewhere.

Activating these properties

Zero credential exposure and multi-hop chain integrity both require architectural changes to take effect. Whereas traditionally this could create complexity and a barrier to deployment, the Outerlimit approach has been specifically streamlined to support enterprise customers.

Firstly, Outerlimit activates an identity-binding policy that ties tool use to the verified identity of the calling agent. The policy declares that the imprint authorizing a given tool call can only reconstruct against the correct composite identity. The same rule applies uniformly across the agents and tools Outerlimit observes. It does not vary per agent or per tool and it does not require workflows to be modelled in advance.

Fine-grained operational policy, for example restricting an agent to contracts below a defined value or limiting a tool to specific teams or time windows, is where the workflow-specific policy enforcement of an agent estate gets expressed. Most enterprise deployments will build there as the estate matures.

The Enterprise Reality

The conditions described in this document do not assume that enterprises are greenfield environments. Instead, they are layered by legacy and new systems, multi-cloud and hybrid topologies and identity that is made complex by people changing roles, retaining historical access and accumulating permissions.

Two challenges stand between an enterprise and effective enforcement:

- **Visibility:** an organization cannot secure what it cannot see. The rapid deployment of uncatalogued and unsanctioned agents, tools and MCP servers means that many enterprises do not hold a complete map of their AI estate.
- **Up to date Policy:** enforcement is only as good as the policy behind it and manual authoring alone cannot keep pace with an estate that changes faster than people can review it.

The Outerlimit Platform delivers a comprehensive pathway to enforcement that starts with discovery and visibility. It integrates with the existing technology stack via APIs to collect data and surface the agentic estate as it exists. Discovered assets are risk-scored by exposure and business impact, producing:

- A prioritized view of the highest-exposure paths.
- A defensible order in which to address them.

The first deliverable is a comprehensive audit of the estate and a sequence for acting on it, before any control is applied.

Enforcement is then brought to paths as they are integrated. Outerlimit generates initial enforcement policies from the execution paths it observes and automatically refines them as that behavior accumulates, so control keeps pace with an ever-changing agentic estate. Each governed path produces attributable records of what was allowed, what was denied, and under which policy, so exposure that was previously invisible becomes constrained and evidenced. Discovery is continuous, with new agents and tools identified over time.

It is important to note that Outerlimit is not an identity remediation tool. The Platform binds enforcement to verified identity, consuming it rather than issuing it. Where an underlying identity is over-permissioned through years of drift, that history is not repaired. What changes however is the management of identity and authority at the point of action. On governed paths there is no standing credential, and authorization is reconstructed for each call against current policy, so accumulated grants no longer translate directly into accumulated executable authority.

Future-proofed Cryptographic Enforcement

As an organization's agent estate grows, the benefits of Outerlimit's cryptographic enforcement compound in the following ways:

Scaling asymmetry

Standing-credential architectures scale multiplicatively. Ten agents calling ten tools on behalf of one hundred users produces ten thousand credential bindings, each one a standing artifact that can be exfiltrated and must be rotated, revoked, and audited. At that scale, maintaining a distinct scoped credential per combination becomes infeasible, so organizations collapse the matrix into shared accounts and broad scopes, resulting in the erosion of least privilege.

Outerlimit's imprint-based architecture scales additively. The same scenario produces approximately one hundred and twenty bindings, because each imprint is meaningful only in the composite. The attack surface and the management burden stay bounded as the estate grows.

Future-proof by design

As industry move toward agent-to-agent architectures, orchestrators, and sub-agents, multi-hop will be the default deployment pattern, not the exception. Outerlimit's multi-hop chain integrity is structural, not procedural, and holds regardless of chain length or topology. As the chain depth of the average deployment grows, the value compounds.

Examination-grade evidence

Auditing standing-credential architectures produces procedural log reports, rotation records, and detection alerts with inferred regulatory control. Audit under Outerlimit's cryptographic enforcement is structural. It produces cryptographic proof that every tool call's authorization was bound to the correct agent identity, including the full chain in multi-agent architectures. Regulators are moving from accepting inferred control toward requiring proof of it. Cryptographic enforcement already produces the proof, and the gap widens as the standard rises.

OWASP Coverage at the Agent Action Layer

The OWASP Top 10 for Agentic Applications (2026) and the OWASP Top 10 for LLM Applications (2025) lists the risks an agent security program must address. The categories operate across multiple layers of the agent security problem. Mapping them against the agent action layer makes the layer structure explicit and shows which categories are addressed by Outerlimit's cryptographic enforcement and which operate at adjacent layers and require additional architectural answers.

OWASP mapped to Outerlimit at the action layer

- Outerlimit's zero credential exposure property addresses the standing-credential surface that drives ASI03 Identity and Privilege Abuse and a portion of LLM02 Sensitive Information Disclosure. Credentials do not exist as standing artifacts, so the architectural conditions for these categories do not hold.

- Outerlimit's multi-hop chain integrity property addresses ASI10 Rogue Agents and a portion of ASI07 Insecure Inter-Agent Communication. A compromised agent in the call chain cannot synthesize the cryptographic contribution of upstream identities and cannot produce a valid credential at the terminal tool call.

- Outerlimit's identity-bound cryptographic enforcement at the agent action boundary addresses ASI02 Tool Misuse and Exploitation and LLM06 Excessive Agency. An agent that attempts to invoke a tool outside its authorized scope cannot reconstruct a valid credential for the call, regardless of whether the attempt originates from goal hijack, prompt injection, or the agent's own decision-making. This is the point at which Least Agency is enforced rather than declared: what an agent is permitted to do is bound to its identity and policy at execution, not only scoped in configuration upstream.

- Outerlimit's identity-bound cryptographic enforcement reduces blast radius for ASI04 Agentic Supply Chain Vulnerabilities. A compromised MCP server, plugin, or third-party tool cannot invoke operations outside the authorized scope of the calling agent, because the credential it would need cannot be reconstructed without the agent's identity contribution.

OWASP mapping at other layers

- Prompt-layer threats, including LLM01 Prompt Injection and ASI01 Agent Goal Hijack, operate at the layer above the agent action layer. An agent that performs a legitimately authorized action against the wrong target, whether manipulated into it or arriving there through its own reasoning, will be permitted by the enforcement point, because the identity and the tool match policy. These categories require architectural answers at the prompt and reasoning layers.

- Memory and context poisoning (ASI06) operate at the agent's internal state layer. Unexpected code execution at the runtime layer (ASI05), cascading failures across agent estates (ASI08), and human-agent trust exploitation (ASI09) operate at adjacent layers and require their own architectural commitments. Model and training-data integrity (LLM03, LLM04) are the work of model-layer vendors.

- The agent security problem is layered. Cryptographic enforcement addresses the agent action layer specifically. Other layers require their own architectural answers, and the architectures that operate at those layers pair with cryptographic enforcement rather than compete with it.

Summary: How to Verify the Architecture in Your Environment

The architectural properties described in this document are available today. Outerlimit is working with an exclusive group of design partners ahead of general availability to verify these architectural solutions in customer environments.

Zero trust transformed enterprise security by moving control from network position to verified identity. It reached the network and identity layers. Agentic AI introduces the agent action layer, where those established controls do not reach. That gap is why agentic AI adoption often stalls in security reviews. Security teams are accountable for what agents do but cannot extend controls to the layer where agents act. As a result, agentic AI programs and deployments have been reduced in scope or cancelled altogether.

Outerlimit’s identity-bound cryptographic enforcement is the architectural commitment that extends zero trust to the agent action layer. With it, agentic AI is no longer an exception to the enterprise's security model. Instead, it is governed by the same principles as everything else the enterprise trusts: verified per request, scoped to least privilege, auditable by construction.

The decision to deploy agentic AI becomes one the security organization can stand behind, rather than a risk it is asked to absorb.

Outerlimit provides a universal security layer for Agentic AI, offering a single control and observability plane across your agentic ecosystem.

As enterprises evolve their adoption of AI and give agents access to tools via APIs, databases and MCP servers, exponential risk is introduced into their environment. Shadow AI and the unpredictability of agents create new attack surfaces and vulnerabilities. To quantify and manage this risk, the ability to discover and control an expanding agentic footprint becomes critical.

Outerlimit closes this visibility and control gap, enabling enterprises to discover and observe their agentic activity, quantify risk, and enforce run-time controls at the tool execution layer.