

The AI Security Gap: A Zero Trust Implementation Guide for Security Teams

A practitioner's guide to securing the full AI lifecycle, from Shadow AI discovery through agentic runtime protection, with zero trust principles, a 30/60/90-day implementation roadmap, and KPIs for leadership reporting.

EBOOK

Introduction

Security teams know they have an AI problem. What most can't answer yet is how big it is.

Which artificial intelligence (AI) apps are running in your environment right now? Which models are your developers querying through integrated development environment (IDE) plugins and command-line interface (CLI) tools? What data is moving through prompts, and is any of its source code, customer personally identifiable information (PII), or content subject to regulatory disclosure rules? Which agents are acting autonomously on behalf of users, and what systems can they reach?

That gap between what AI is doing in your environment and what you can see and govern is where risk actually lives.

Not in theoretical attack scenarios, but in the daily, unsanctioned, well-intentioned use of AI tools that security never approved because no one thought to ask. Shadow AI and embedded AI features inside already-approved software as a service (SaaS) platforms are the primary sources of the gap. A tool that IT approved for one purpose shipped an AI assistant by default, and now it's processing internal data without any security review.

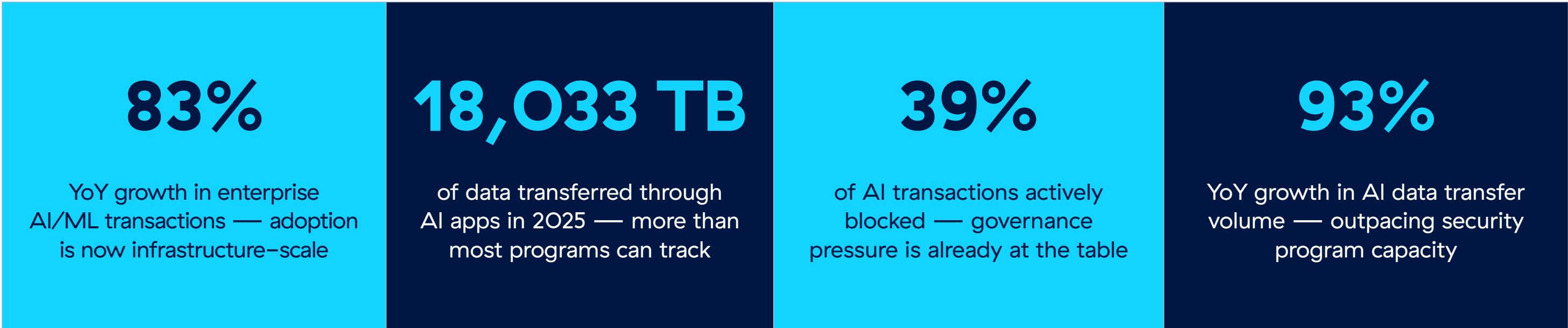
Data and statistics in this guide are drawn from the ThreatLabz 2026 AI Security Report, published by Zscaler in 2026, based on analysis of 989.3 billion AI/ML transactions observed across the Zscaler Zero Trust Exchange from January—December 2025. All figures reflect that research period unless otherwise noted.





Zero trust offers a principled response:

Treat no app, user, prompt, or connector as inherently trusted. Enforce identity- and context-based access at every layer. Verify continuously. Minimize blast radius when something goes wrong. The challenge is sequencing. Visibility has to come before policy enforcement. Access controls have to be stable before automation is layered on top. Build out of sequence and you produce controls that either block legitimate work or fail to stop actual risk.



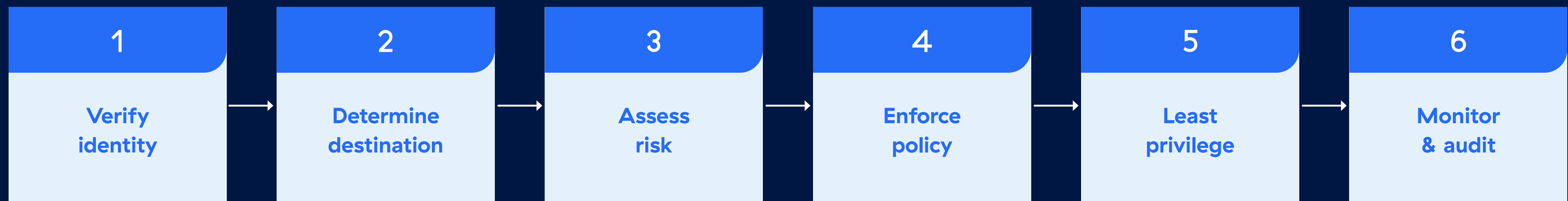
Source: ThreatLabz 2026 AI Security Report, Zscaler (pp. 5—6)

What this guide covers

- 1. The seven zero trust principles applied specifically to AI systems, agents, and data flows
- 2. A phased 30/60/90-day implementation roadmap built in dependency order
- 3. Coverage of emerging AI risk domains: shadow AI, prompt injection, agentic systems, Model Context Protocol (MCP)/agent-to-agent (A2A) traffic, embedded SaaS AI, and AI infrastructure posture
- 4. A KPI model for reporting AI security program progress to leadership

Zero trust principles applied to AI

The seven principles below form the decision framework for everything in the implementation roadmap. Each one is a decision-making tool for evaluating controls. The sequencing mirrors how a security team would actually apply them in practice.



Applied in sequence. Each principle builds on the one before it.

Principle 1:

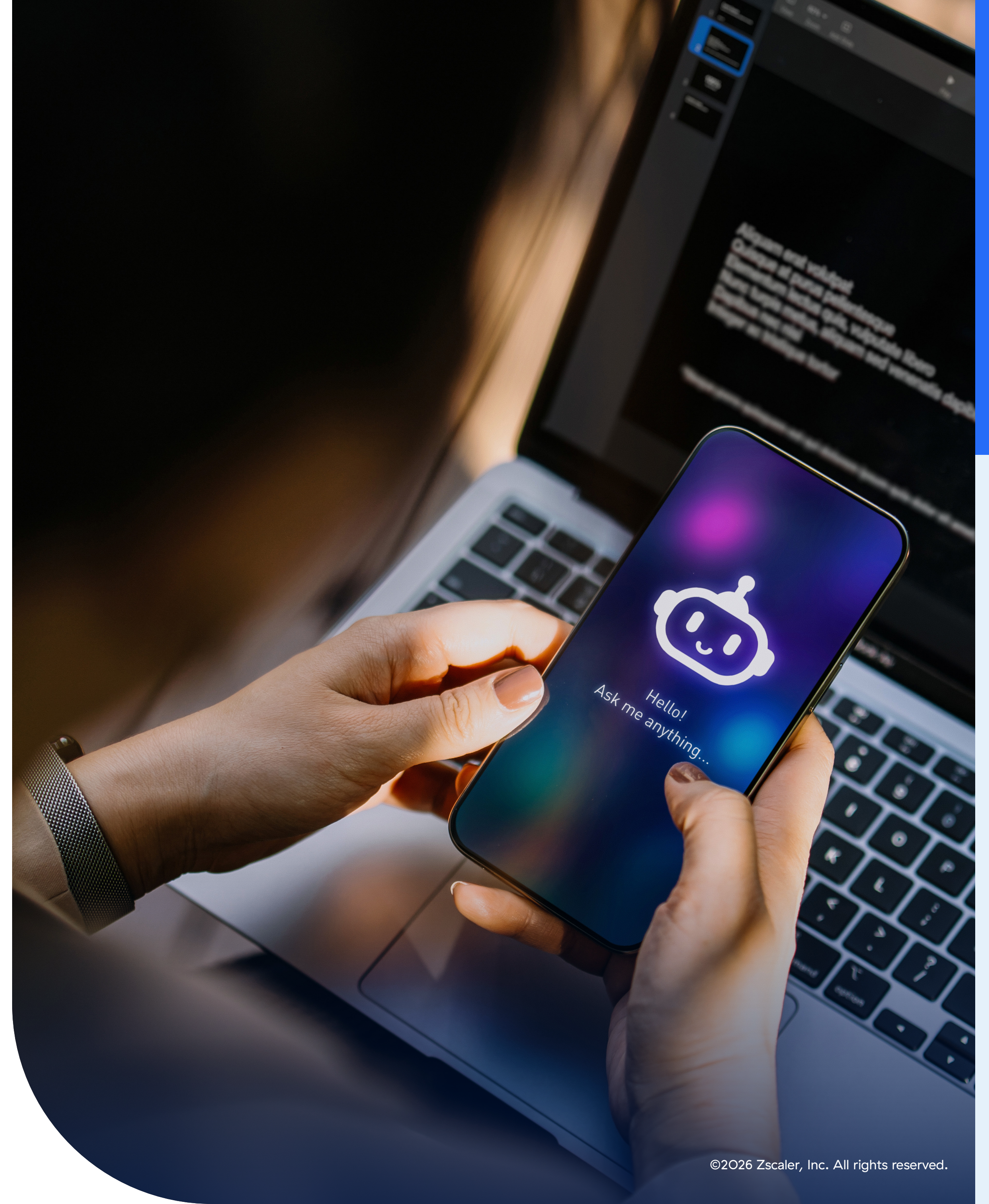
VERIFY IDENTITY FOR USERS, DEVELOPERS, AND SERVICES

AI access begins with identity. Every user accessing a generative AI (GenAI) SaaS tool, every developer querying a model endpoint from an IDE plugin, and every service account connecting to an AI infrastructure component should authenticate through a consistent identity control plane. Single sign-on (SSO) and multifactor authentication (MFA) are baseline, but the details matter.

Workforce and developer populations carry different risk profiles and require distinct policy treatment. A marketing analyst using ChatGPT for content drafts presents different data exposure risks than a software engineer whose IDE is routing code completions through a GitHub Copilot plugin connected to an MCP server with access to internal repositories. Service-to-service authentication for AI agents is the emerging edge case most organizations are not ready for.

Key controls

- SSO and MFA enforcement for AI SaaS and AI developer tooling
- Separate identity policies for workforce GenAI usage versus developer AI tooling (IDEs, CLIs, MCP integrations)
- **Service account controls for AI agents:** least-privilege scoping, short-lived credentials, and audit logging
- **Device posture as an identity signal:** managed endpoints get different access than bring your own device (BYOD)



Principle 2:

DETERMINE DESTINATION FOR SANCTIONED AND UNSANCTIONED AI

Not all AI apps are equal, and a blanket block-everything approach is neither operationally realistic nor security-effective. **The goal is a tiered destination policy:** sanctioned tools get full access with inline inspection, tolerated tools get monitored access with usage guardrails, and unsanctioned or high-risk tools get blocked or isolated. That list of AI surfaces is larger than most organizations expect, growing fourfold in a single year as new AI capabilities ship inside existing platforms.

Key controls

- Tiered destination policies to allow, warn, isolate, or block by app and user group
- Dynamic app discovery to surface new AI tools as they appear in traffic
- Separate policy tracks for GenAI SaaS, embedded AI in enterprise SaaS, and AI developer tooling
- Approval workflow for new AI tool requests so users have a path forward rather than working around controls

3,400+

**DISTINCT AI APPLICATIONS TRACKED ACROSS ENTERPRISE TRAFFIC
4X GROWTH IN A SINGLE YEAR, AND RISING MONTHLY**

ThreatLabz 2026 AI Security Report, Zscaler (p. 5)



Principle 3:

ASSESS RISK WITH CONTEXT-BASED DECISIONS

Access decisions for AI should never be binary. The same user accessing the same AI tool presents different risks depending on context: device, data sensitivity, and action. Effective risk assessment requires defining red data upfront, source code, PII, Payment Card Industry (PCI) data, protected health information (PHI), unreleased financial results, and confidential merger and acquisition (M&A) details. The top violation categories surfaced in enterprise traffic tell you exactly what employees are submitting: name and national identifier data, source code, and medical records.

Key controls

- Explicit “red data” categories for source code, PII/PCI/PHI, secrets, and unreleased plans
- Risk scoring by user group, department, device posture, and data sensitivity
- Dynamic policy adjustment based on contextual signals, not just static allow/block rules
- Separate risk treatment for AI interactions that include file uploads versus text-only prompts

410M+

DATA LOSS PREVENTION (DLP) VIOLATIONS LINKED TO CHATGPT ALONE
TOP TYPES: NAME LEAKAGE, SSNS, SOURCE CODE, MEDICAL DATA

ThreatLabz 2026 AI Security Report, Zscaler (pp. 6, 15)



Principle 4:

ENFORCE POLICY INLINE, PER SESSION

Logs tell you what happened. By the time you find a record showing a developer submitted proprietary source code to an unsanctioned model, the data is already gone. Most organizations compound this by deploying DLP for file uploads to AI tools while leaving prompt text uninspected. A developer typing source code directly into a prompt window clears those controls without triggering a single alert. The data transfer growth shows the scale of what is moving through those unguarded channels.

Key controls

- Inline DLP on prompts, responses, and file uploads, not just at the URL/category level
- Allow, warn, redact, block, and isolate actions available at policy enforcement points
- Prompt classification for leading GenAI apps with context-aware detection
- Content moderation for off-topic, restricted, and competitive-sensitive interactions
- Browser isolation for allowed-but-risky AI tools: access without full trust

93%

**YOY GROWTH IN ENTERPRISE AI DATA TRANSFER VOLUME OUTPACING
WHAT MOST SECURITY PROGRAMS CAN TRACK OR GOVERN**

ThreatLabz 2026 AI Security Report, Zscaler (pp. 5, 14)

Principle 5:

MINIMIZE ACCESS AND EXPOSURE

AI systems create access paths that often sit outside traditional identity and network models. A Retrieval-Augmented Generation (RAG) pipeline may give a model access to a document corpus. An agent may connect to internal tools, datasets, vector stores, or MCP servers. Those permissions are as real as any user's access, even if they do not appear in standard access reviews.

Minimizing access and exposure means reducing what AI systems can reach and constraining how AI components communicate. Agents, data pipelines, RAG components, and backend data sources should only connect through explicitly approved paths. Without least-privilege scoping and segmentation, an overprivileged agent or compromised AI component can become a lateral movement path.

Key controls

- Role-based access scoping for AI apps, agents, model connectors, datasets, vector stores, and AI infrastructure components
- Regular review and reduction of overprivileged access to sensitive datasets and AI-connected systems
- Minimal permission sets for AI agents: define what they can access, restrict everything else
- Segmentation of AI services, agent platforms, data pipelines, RAG components, and backend data sources
- Restrict east-west access between AI components to explicitly defined paths
- Protocol-aware inspection and logging for MCP and A2A traffic

Minimize blast radius at every AI access path

Treat every AI connector, agent permission, RAG pipeline, vector store, MCP server, and component-to-component connection as an exposure point. Scope what it can access, segment where it can communicate, inspect the traffic, and log the activity.



Principle 6:

CONTINUOUS MONITORING AND AUDITABILITY

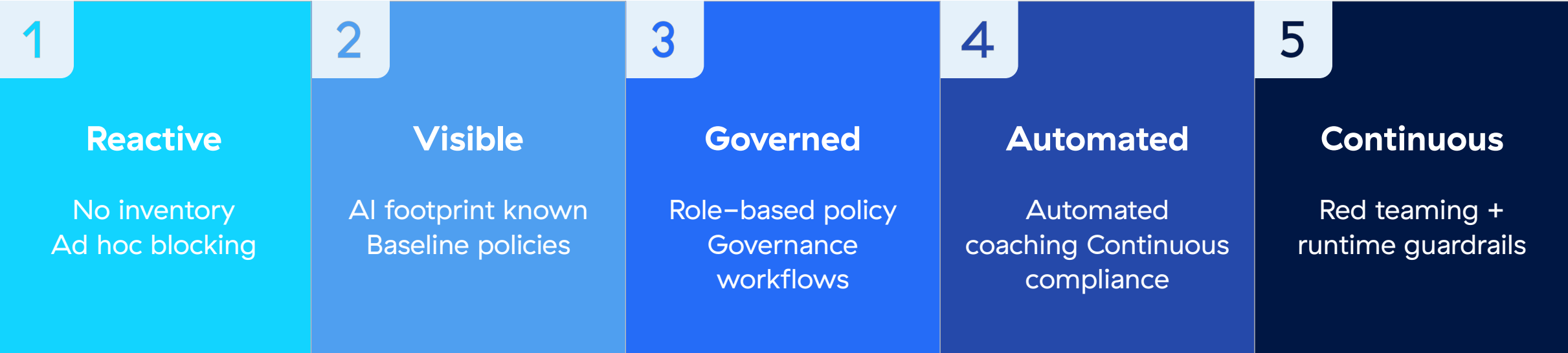
Continuous verification is the operating model, not a milestone you reach and move past. For AI systems that means prompt and response logging, real-time alerting on policy violations, usage dashboards, and an investigation workflow that does not require manual log correlation across five different tools.

The regulatory pressure behind this is real: the [European Union Artificial Intelligence Act \(EU AI Act\)](#), [National Institute of Standards and Technology Artificial Intelligence Risk Management Framework \(NIST AI RMF\)](#), and emerging AI governance frameworks all require demonstrable, documented controls.

Key controls

- Prompt and response logging for inline AI interactions
- Usage dashboards by user, department, and AI app with risk signal overlays
- Alerting and investigation workflow for AI-related policy violations and anomalies
- Audit-ready reporting: what AI is in use, what controls are applied, and what violations occurred
- Governance posture mapping to NIST AI RMF, EU AI Act, and other applicable frameworks





Program maturity →

Day 1–30
Establish Visibility

You cannot secure AI you cannot see. The first 30 days focus on building enough inventory and baseline control to make informed decisions, not on locking everything down

DAY 1–30 Establish visibility	DAY 31–60 Build policy maturity	DAY 61–90 Automate and validate
• Deploy AI app discovery • Define red data categories • Baseline access policies • First leadership dashboard	• Role-based access policies • Governance workflows live • AI-SPM rollout • Private AI protections	• Automated user coaching • Continuous compliance • Red teaming cadence • Closed-loop guardrails

30/60/90-day implementation roadmap at a glance



30-day implementation checklist

- ✓ **Establish a recurring AI tool review cadence:** new tools discovered in traffic, exception requests received, and policy changes made
- ✓ Build the first leadership-facing AI security dashboard: AI app footprint (sanctioned vs. shadow), DLP events by category, policy actions, and trend lines
- ✓ Define incident playbooks for the most common AI-related security events: sensitive data in prompts, access to blocked destinations, large file uploads to AI tools
- ✓ **Centralize AI security logs:** user activity, app usage, prompt/response events, policy actions, and violations in a single dashboard view

Operationalize

- ✓ **Review and tune DLP policies using the first two weeks of telemetry:** reduce false positives, improve detection on confirmed data exposure patterns
- ✓ **Enable content moderation for acceptable-use enforcement:** off-topic interactions, restricted topics, competitive-sensitive content, and inappropriate outputs
- ✓ **Add browser isolation for AI tools in the allowed-but-risky tier:** users can access the tool but uploads, downloads, and copy/paste are restricted at the browser layer
- ✓ **Expand DLP coverage to full red data scope:** source code (additional languages and patterns), PII (beyond obvious formats), PCI cardholder data, PHI, and confidential business content

Expand data protection

- ✓ **Publish an AI acceptable-use traffic light guide:** green (sanctioned), yellow (use with restrictions), red (blocked, with a reason). Make the exception request process visible
- ✓ Enable inline DLP for prompts, starting with high-confidence detection rules for red data categories: source code patterns, credential formats, PII schemas

- ✓ **Restrict the highest-risk actions on unsanctioned tools:** uploads, large copy/paste blocks, and download of AI outputs to unmanaged storage
- ✓ **Enforce destination policy:** warn or block by user group and app for clearly unsanctioned or high-risk AI destinations

Apply baseline controls

- ✓ **Establish a baseline AI risk inventory:** which apps are accessing which data types, and which user groups are the highest-volume AI users
- ✓ **Define red data categories explicitly:** source code, PII, PCI data, PHI, secrets and credentials, unreleased financial data, confidential M&A or strategic plans

Define what needs protecting

- ✓ Enable prompt visibility for highest-traffic AI apps to understand what data is moving through prompts and responses
- ✓ Map which AI tools have integrations with internal systems and what data types they access
- ✓ **Surface shadow AI:** tools in active use that were never formally approved
- ✓ Deploy AI app discovery to inventory GenAI SaaS, embedded AI in enterprise platforms, and AI developer tooling in use across the environment

TRADE-OFF TO CONSIDER

Starting with warning policies rather than hard blocks reduces friction and surfaces false positives in your DLP rules before they start blocking legitimate work. The cost is that data exposure events can still occur during the warn phase. Balance this against your red data risk tolerance, if source code leakage is a critical risk, hard blocks on code-pattern detection are worth the early false-positive pain.



30-day deliverables

- AI app inventory with risk tier classification (sanctioned, tolerated, prohibited)
- Initial access policies live for highest-risk AI apps and user groups
- Prompt DLP active for red data categories with initial tuning complete
- First leadership dashboard with AI footprint and risk metrics
- AI acceptable-use guide and exception request process published

Day 31–60

Build policy maturity

Days 31–60 build on visibility and initial controls to establish a mature policy model, functioning governance workflows, and AI infrastructure inventory.



60-day implementation checklist

Mature the access policy model

- ✓ **Build department-specific AI access policies:** HR, Finance, Engineering, and Sales have different AI tool needs and data exposure risk profiles
- ✓ **Layer conditional policy controls:** managed device versus BYOD access tiers, geographic access signals, device posture risk scoring
- ✓ **Formalize three access tiers for AI apps:** sanctioned (full access with inline inspection), tolerated (monitored access with usage restrictions), prohibited (blocked with exception path)
- ✓ Review and sunset exception requests from the initial 30 days against the now-more-mature policy model

Establish governance processes

- ✓ **Map AI governance requirements to internal policies and applicable frameworks:** NIST AI RMF, EU AI Act, HIPAA where PHI is in scope
- ✓ **Formalize the AI tool approval workflow:** submission, security review, risk classification, conditional approval, and periodic revalidation
- ✓ Establish standard sensitivity labels for data types relevant to AI usage, with clear handling rules for each label
- ✓ **Define escalation paths for AI security incidents:** who owns response, what constitutes an incident requiring executive notification, and what evidence is preserved

Inventory AI infrastructure

- ✓ Inventory AI models, agents, services, datasets, vector stores, and pipelines, and build the AI-BOM with lineage from datasets through to runtime usage
- ✓ **Run AI-SPM assessment:** identify misconfigurations, excessive permissions, sensitive data exposure, and vulnerability findings across AI infrastructure components
- ✓ **Prioritize high-risk AI-SPM remediation:** excessive permissions, exposed sensitive data, and unconstrained AI component access paths
- ✓ **Validate coverage across major AI platforms:** cloud-hosted models, enterprise-deployed agent platforms, MCP server integrations, and unmanaged AI infrastructure

Protect private AI applications

- ✓ **Add prompt injection detection to controls for private AI apps:** adversarially crafted inputs designed to override model instructions or extract sensitive information
- ✓ **Implement data poisoning monitoring for RAG pipelines:** detect unauthorized modifications to the document corpus that could skew model outputs
- ✓ **Enable prompt/response logging for private AI systems with anomaly detection:** unusual query patterns, unexpected output content, and off-scope interactions
- ✓ Establish model interaction review workflows for high-sensitivity private AI applications

60-day deliverables

- Mature role-based access policy model with conditional enforcement live
- AI governance workflow operational: tool approval, sensitivity labeling, escalation paths defined
- AI-SPM visibility baseline complete with prioritized remediation backlog
- AI-BOM covering major AI infrastructure components
- Private AI protections in place for highest-priority internal AI apps
- Compliance reporting cadence defined and first report produced

Day 61–90

Automate
and validate

Days 61–90 make the program self-sustaining: automated enforcement, continuous compliance monitoring, and adversarial validation through red teaming.



90-day implementation checklist

Automate enforcement

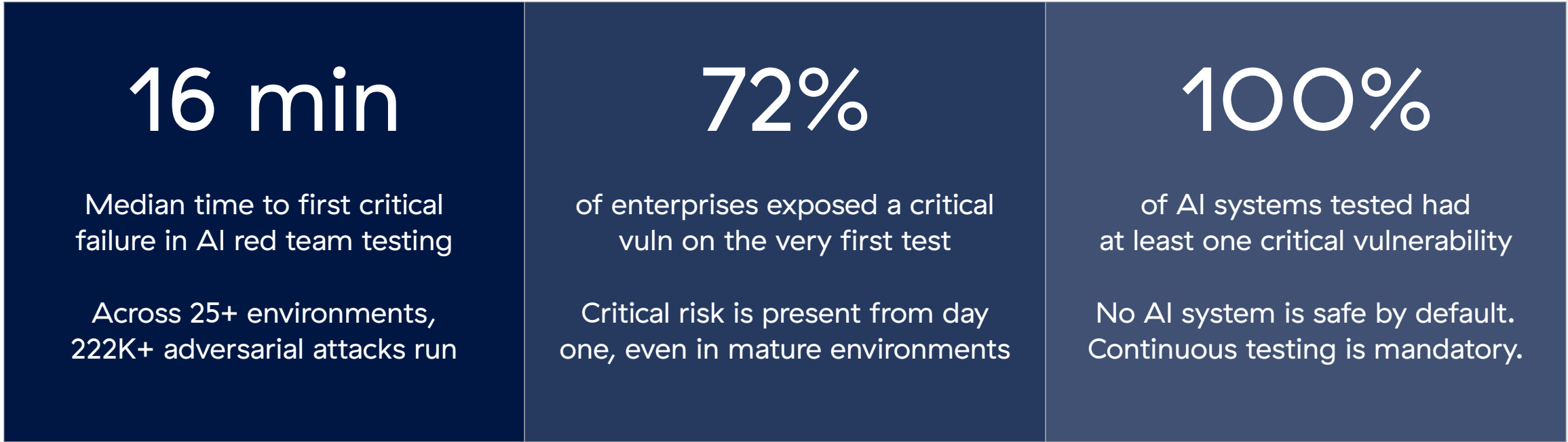
- ✓ **Implement automated user coaching for policy violations:** warn with context-specific guidance before escalating to block
- ✓ **Integrate ITSM or SOAR ticketing for repeated violations:** create investigation tickets automatically for users who repeatedly trigger the same controls
- ✓ **Configure auto-quarantine and isolation triggers for high-risk behaviors:** large-volume sensitive data submissions, access to prohibited AI destinations, anomalous prompt patterns
- ✓ **Build feedback loops:** track policy violation trends over time to identify categories of behavior that suggest policy tuning needs rather than enforcement failures

Lock in compliance reporting

- ✓ **Configure always-on monitoring for AI compliance posture:** which controls are active, which assets are covered, and where coverage gaps exist
- ✓ **Map controls to applicable frameworks:** NIST AI RMF (including NIST AI 600-1 for generative AI risk), EU AI Act obligations, GDPR, and HIPAA where PHI is in scope
- ✓ **Build audit-ready reporting packages:** AI app usage summary, prompt DLP actions and outcomes, policy exceptions granted and expired, AI-SPM finding remediation status
- ✓ **Establish a quarterly AI security program review cadence:** policy performance, coverage gaps, new risk areas, and roadmap adjustments

Red team and harden

- ✓ **Establish AI red teaming cadence for internally built AI applications:** prebuilt attack probe libraries plus customized test cases targeting your specific application context
- ✓ **Run tests covering the full AI attack surface:** prompt injection, jailbreaks, context poisoning, PII extraction, source code leakage, off-topic behavior, and model benchmark regression
- ✓ **Deploy runtime protection policies for production AI systems:** detectors for prompt injection, data leakage (PII and source code), jailbreak attempts, and content moderation violations
- ✓ **Automate the translation of red teaming findings into runtime guardrails:** the gap between testing and production enforcement is where most AI application security programs break down





Validate architecture alignment

- ✓ **Map all AI access paths against the six zero trust principles:** verify each principle is applied across workforce, developer, and AI infrastructure access surfaces
- ✓ **Identify and close policy gaps at the intersections:** places where one enforcement layer ends and another begins without consistent coverage
- ✓ Standardize enforcement through a unified policy model spanning identity verification, destination control, risk assessment, inline enforcement, access and exposure minimization, and monitoring

90-day deliverables

- Automated violation response and user coaching workflows operational
- Continuous compliance monitoring live with framework alignment documented
- Governance reporting package for leadership and audit use
- AI red teaming cadence established with runtime protection policies active
- Closed-loop translation of red teaming findings to production guardrails
- Program scorecard and quarterly roadmap for ongoing governance

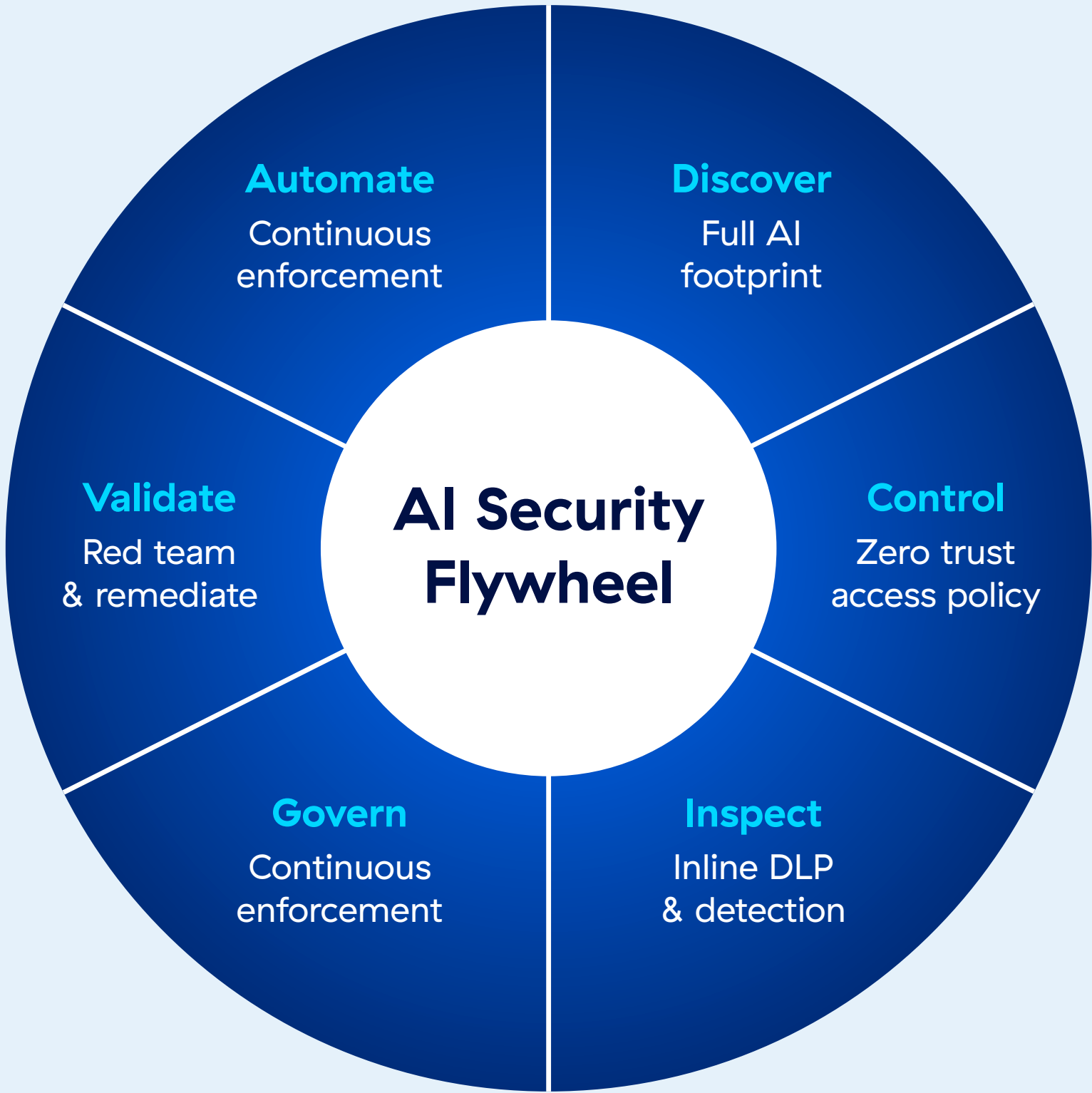


How Zscaler enables secure AI adoption

The Zero Trust Exchange™ platform was built to govern and secure AI at enterprise scale. It extends the same platform that already governs user, workload, and network access for thousands of enterprises globally, and is not retrofitted onto a legacy security architecture.

Each stage in the flywheel maps to a specific set of platform capabilities. The table below shows what Zscaler delivers at each layer.

DISCOVER Know your full AI footprint	CONTROL Enforce zero trust access for AI	PROTECT Harden and defend AI applications
AI Asset Management Zscaler accesses every AI application, model, agent, MCP server, and embedded SaaS AI feature across your environment, including tools IT never approved.	AI Access Security Enforce zero trust access controls for AI SaaS, embedded AI in enterprise platforms, and AI developer environments with inline inspection at every layer.	AI Red Teaming & AI Guardrails Test privately built AI applications against real adversarial conditions, then automatically translate findings into runtime protection policies.
- Shadow AI discovery and usage visibility across all user populations AI Bill of Materials (AI-BOM): lineage from datasets through models, agents, and runtime usage AI-SPM: continuous posture assessment for misconfigurations, excessive permissions, and sensitive data exposure - AI agent detection across enterprise-deployed and SaaS-embedded agents	- Granular allow, warn, isolate, and block policies by user group, app, and data type - Inline DLP on prompts, responses, and file uploads across all AI traffic - Zero Trust Browser for unmanaged and BYOD access to AI tools - Prompt extraction and classification for leading GenAI apps	Automated, continuous red teaming: prompt injection, jailbreaks, context poisoning, data leakage, behavior drift - Prompt hardening and model benchmarking with governance mapping to NIST AI RMF, EU AI Act, OWASP LLM Top 10 Runtime guardrails: detectors for jailbreaks, PII/source code leakage, content moderation, and off-topic behaviors Closed-loop policy generation: red teaming findings become production detectors automatically



THE PLATFORM ADVANTAGE

Most security vendors solve one slice of the AI security problem. Zscaler covers the full AI lifecycle on a single platform: discover AI everywhere, control who accesses what AI using zero trust policy, and protect AI apps and infrastructure from build through runtime. The closed loop between red teaming findings and production runtime guardrails is a capability no point solution can replicate.

Frequently asked questions

What does ‘zero trust AI’ mean in practice?

Zero trust AI means applying the core zero trust principle, assume no implicit trust, verify continuously, to every AI interaction in your environment. In practice: no AI app, user, device, prompt, or connector is inherently trusted. Access is granted based on verified identity, device posture, data sensitivity, and session context. Controls are applied inline, in real time. Every access decision is logged and auditable.

What should we do in the first 30 days to reduce shadow AI risk?

Start with discovery. You cannot govern AI you cannot see, and ThreatLabz tracks more than 3,400 AI applications in enterprise traffic, a number that has grown fourfold in a single year, driven largely by embedded AI features shipping inside tools IT already approved. Once you have visibility, classify the highest-risk apps and set baseline destination policies. Add prompt DLP for your highest-risk data categories. Publish a transparent acceptable-use guide and exception request process so users have a compliant path forward rather than incentive to work around controls.

How do AI access controls and inline DLP protect prompts and uploads?

Access controls determine who can reach which AI tools, from which devices and locations, under what conditions. Inline DLP operates at the content layer, inspecting what users actually submit, prompt text, file uploads, and copy/paste interactions, in real time. The two controls are complementary: access control limits who reaches which AI, DLP controls what data moves through those permitted channels. A gap in either creates exposure, which is why organizations that deploy DLP for uploads but skip prompt inspection still carry significant leakage risk.

What is AI-SPM, and why does it matter for AI adoption?

AI-SPM is continuous assessment of AI infrastructure configurations, permissions, and exposure risks, the AI-equivalent of CSPM for cloud infrastructure. It discovers AI assets, evaluates their configurations against security best practices, identifies excessive permissions and sensitive data exposure, and tracks posture drift over time. It matters because the misconfiguration surface area in AI infrastructure is large and poorly understood by most organizations moving from AI experimentation to deployment.

What KPIs best demonstrate to leadership that AI risk is decreasing without blocking productivity?

The most effective framing pairs a risk reduction metric with an adoption metric so leadership sees both trends together. Shadow AI app count decreasing alongside sanctioned AI adoption increasing is a clear signal that governance is working. Prompt DLP actions (block versus warn versus allow distribution) show how enforcement is calibrated. Exception request turnaround time and AI tool approval cycle time show governance responsiveness.

ZSCALER AI SECURITY

Zero Trust Exchange Platform

AI Asset Management

AI Access Security

AI Red Teaming

AI Guardrails

zscaler.com/ai-security





Act Fast. Stay Secure.

About Zscaler

Zscaler (NASDAQ: ZS) is a pioneer and global leader in zero trust security. The world's largest businesses, critical infrastructure organizations, and government agencies rely on Zscaler to secure users, branches, applications, data & devices, and to accelerate digital transformation initiatives. Distributed across more than 160 data centers globally, the Zscaler Zero Trust Exchange™ platform combined with advanced AI combats billions of cyber threats and policy violations every day and unlocks productivity gains for modern enterprises by reducing costs and complexity.

© 2026 Zscaler, Inc. All rights reserved. Zscaler™ and other trademarks listed at [zscaler.com/legal/trademarks](https://www.zscaler.com/legal/trademarks) are either (i) registered trademarks or service marks or (ii) trademarks or service marks of Zscaler, Inc. in the United States and/or other countries. Any other trademarks are the properties of their respective owners.

+1 408.533.0288

Zscaler, Inc. (HQ) • 120 Holger Way • San Jose, CA 95134

[zscaler.com](https://www.zscaler.com)