Die KI-Sicherheitslücke: Leitfaden zur Zero-Trust-Implementierung für Sicherheitsexperten

Ein Leitfaden für Praktiker zur Absicherung des gesamten KI-Lebenszyklus: von der Erkennung von Schatten-KI bis hin zum agentenbasierten Laufzeitschutz, basierend auf Zero-Trust-Prinzipien, Umsetzungsplan (30/60/90 Tage) und KPIs für das Management-Reporting.

E-BOOK

Einführung

Sicherheitsexperten wissen, dass sie ein KI-Problem haben. Die meisten können nur seine Größe noch nicht abschätzen.

Welche Anwendungen künstlicher Intelligenz (KI) laufen derzeit in Ihrer Umgebung? Welche Modelle fragen Ihre Entwickler über IDE-Plugins (Integrated Development Environment) und CLI-Tools (Command-Line Interface) ab? Welche Daten werden durch die Prompts übertragen? Unterliegen der Quellcode, die personenbezogenen Daten der Kunden oder der Inhalt aufsichtsrechtlichen Offenlegungsvorschriften? Welche Agents agieren autonom im Auftrag der Nutzer und auf welche Systeme können sie zugreifen?

Die Diskrepanz zwischen dem tatsächlichen Verhalten von KI in Ihrer Umgebung und dem, was Sie sehen und steuern können, birgt das größte Risiko. Nicht in theoretischen Angriffsszenarien, sondern im alltäglichen, nicht genehmigten und gut gemeinten Einsatz von KI-Tools, die nie von der IT-Sicherheit freigegeben wurden, weil niemand nachfragte. Schatten-KI und eingebettete KI-Funktionen in bereits genehmigten SaaS-Plattformen (Software as a Service) sind die Hauptursachen dieser Diskrepanz. Ein Tool, das die IT für einen bestimmten Zweck freigegeben hat, enthielt standardmäßig einen KI-Assistenten und verarbeitet nun interne Daten ohne jegliche Sicherheitsprüfung.

Die Daten und Statistiken in diesem Leitfaden stammen aus dem ThreatLabz-Report zur KI-Sicherheit, der 2026 von Zscaler veröffentlicht wurde. Der Report basiert auf der Analyse von 989,3 Milliarden KI/ML-Transaktionen, die im Zeitraum Januar bis Dezember 2025 über die Zscaler Zero Trust Exchange verarbeitet wurden. Alle Zahlen beziehen sich auf den Untersuchungszeitraum, sofern nicht anders angegeben.

Zero Trust ermöglicht eine prinzipienbasierte Lösung:

Keine App, kein User, kein Prompt und kein Konnektor darf als grundsätzlich vertrauenswürdig betrachtet werden. Durchsetzung identitäts- und kontextbasierter Zugriffskontrollen auf allen Ebenen. Kontinuierliche Überprüfung. Minimierung des Schadenspotenzials im Ernstfall. Die Herausforderung liegt in der richtigen Abfolge. Transparenz muss vor der Durchsetzung von Richtlinien kommen. Zugriffskontrollen müssen stabil sein, bevor die Automatisierung darauf aufbauen kann. Wenn die Kontrollmechanismen nicht in der richtigen Reihenfolge implementiert werden, entstehen entweder Kontrollen, die legitime Arbeitsabläufe blockieren, oder solche, die tatsächliche Risiken nicht abwenden.

83 %	**18.033 TB**	**39 %**	**93 %**
Zunahme der KI/ML-Transaktionen im Vergleich zum Vorjahr — die Implementierung hat nun einen Infrastruktur-Maßstab erreicht.	Daten wurden im Jahr 2025 über KI-Apps übertragen — mehr, als die meisten Programme erfassen können.	KI-Transaktionen wurden aktiv blockiert — der Governance-Druck ist bereits spürbar.	Wachstum des KI-Datentransfervolumens im Vergleich zum Vorjahr — übertrifft die Kapazität von Sicherheitsprogrammen

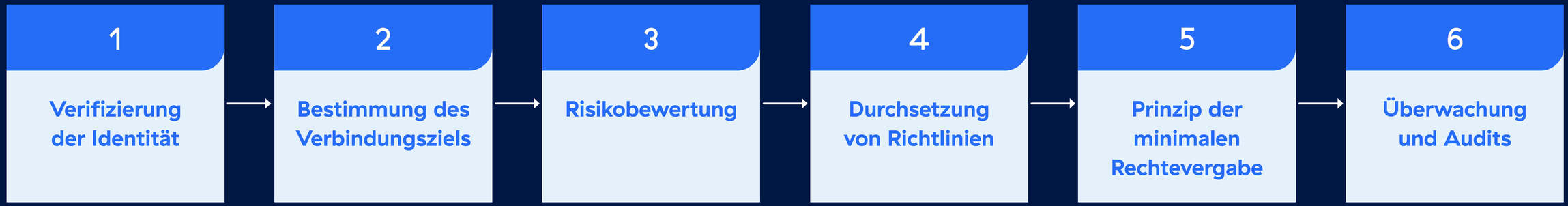
Quelle: Report zur KI-Sicherheit von Zscaler Threat Labz (S. 5—6)

Schwerpunkte des Leitfadens

1. Konkrete Anwendung der sieben Zero-Trust-Prinzipien auf KI-Systeme, -Agents und -Datenflüsse
2. Implementierungsplan für 30/60/90 Tage mit integrierter Abhängigkeitsreihenfolge
3. Empfehlungen für neu entstehende KI-Risikobereiche: Schatten-KI, Prompt-Injection, agentenbasierte Systeme, Modellkontextprotokoll (MCP)/Agent-to-Agent (A2A)-Traffic, eingebettete SaaS-KI, Sicherheitsstatus von KI-Infrastrukturen
4. KPI-Modell zur Berichterstattung über den Fortschritt von KI-Sicherheitsprogrammen an die Führungsebene

Anwendung von Zero-Trust-Prinzipien

Die folgenden sieben Prinzipien bilden den Entscheidungsrahmen für alle Elemente des Implementierungsplans. Jedes dieser Instrumente dient der Entscheidungsfindung und Bewertung von Kontrollmaßnahmen. Die Abfolge spiegelt wider, wie ein Sicherheitsteam sie in der Praxis anwenden würde.



Sequenzielle Umsetzung. Jedes Prinzip baut auf dem vorhergehenden auf.

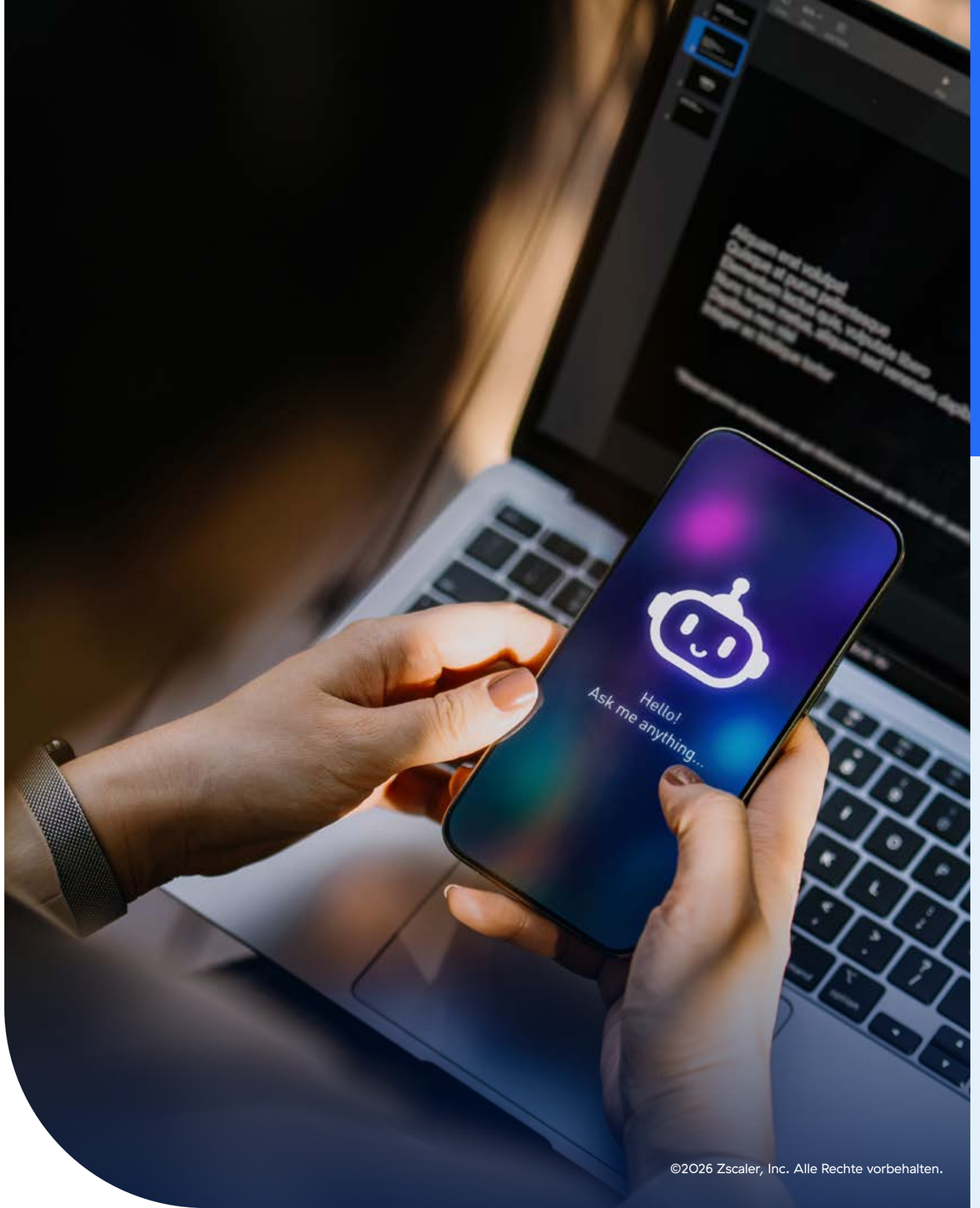
1. Grundsatz:
IDENTITÄT VON USERN, ENTWICKLERN UND SERVICES ÜBERPRÜFEN

Der Zugang zu KI beginnt mit der Identität. Jeder User, der auf ein generatives KI-SaaS-Tool (GenAI) zugreift, jeder Entwickler, der einen Modellendpunkt von einem IDE-Plugin aus abfragt, und jedes Servicekonto, das eine Verbindung zu einer KI-Infrastrukturkomponente herstellt, sollte sich über eine konsistente Identitätskontrollebene authentifizieren. Single Sign-On (SSO) und Multifaktor-Authentifizierung (MFA) sind Standard, aber es kommt auf die Details an.

Arbeitskräfte und Entwickler weisen unterschiedliche Risikoprofile auf und erfordern daher eine differenzierte Behandlung. Ein Marketinganalyst, der ChatGPT für Content-Entwürfe verwendet, birgt andere Risiken der Datenoffenlegung als ein Softwareentwickler, dessen IDE Codevervollständigungen über ein GitHub Copilot-Plugin weiterleitet, das mit einem MCP-Server verbunden ist und Zugriff auf interne Repositories hat. Die Service-zu-Service-Authentifizierung für KI-Agents ist der aufkommende Grenzfall, auf den die meisten Unternehmen nicht vorbereitet sind.

Die wichtigsten Kontrollmaßnahmen

- SSO- und MFA-Durchsetzung für KI-SaaS und KI-Entwicklertools
- Separate Identitätsrichtlinien für die Nutzung von GenAI durch die Belegschaft im Vergleich zu KI-Tools für Entwickler (IDEs, CLIs, MCP-Integrationen)
- **Kontrolle von Servicekonten für KI-Agents:** minimale Rechtevergabe, kurzlebige Anmeldedaten und Audit-Protokollierung
- **Gerätestatus als Identitätssignal:** Verwaltete Endgeräte erhalten andere Zugriffsrechte als unverwaltete Endgeräte von Usern.



2. Grundsatz:
BESTIMMUNG DES ZIELS FÜR GENEHMIGTE UND INOFFIZIELL GENUTZTE KI

Nicht alle KI-Anwendungen sind gleich, und ein pauschaler Blockierungsansatz ist weder betrieblich realistisch noch sicherheitseffektiv. **Ziel ist eine gestaffelte Zugriffsrichtlinie:** Zugelassene Tools erhalten vollen Zugriff mit integrierter Prüfung, tolerierte Tools erhalten überwachten Zugriff mit Nutzungsbeschränkungen, und nicht zugelassene oder risikoreiche Tools werden blockiert oder isoliert. Die Liste der KI-Anwendungen ist umfangreicher als von den meisten Organisationen erwartet und vervierfacht sich innerhalb eines Jahres, da neue KI-Funktionen in bestehende Plattformen integriert werden.

Die wichtigsten Kontrollmaßnahmen

- Gestaffelte Zielrichtlinien zum Zulassen, Warnen, Isolieren oder Blockieren nach App und Usergruppe
- Dynamische App-Erkennung zur Anzeige neuer KI-Tools, sobald diese im Traffic auftauchen.
- Separate Richtlinien für GenAI SaaS, eingebettete KI in Enterprise SaaS und KI-Entwicklertools
- Genehmigungsworkflow für neue KI-Tool-Anfragen, damit User einen klaren Weg vorfinden, anstatt Kontrollen umgehen zu müssen.

>3.400

DIE ANZAHL DER IM GESAMTEN UNTERNEHMENSTRAFFICS ERFASSTEN KI-ANWENDUNGEN HAT SICH INNERHALB EINES JAHRES VERVIERFACHT UND STEIGT MONATLICH WEITER AN.

Report von Zscaler ThreatLabz zur KI-Sicherheit 2026 (S. 5)

3. Grundsatz:
RISIKOBEWERTUNG ANHAND KONTEXTBEZOGENER ENTSCHEIDUNGEN

Zugriffsentscheidungen für KI dürfen niemals binär sein. Der Zugriff desselben Nutzers auf dasselbe KI-Tool birgt je nach Kontext unterschiedliche Risiken: Gerät, Datensensibilität und Aktion. Eine effektive Risikobewertung erfordert die frühzeitige Definition kritischer Daten, darunter Quellcode, personenbezogene Daten, PCI-Daten (Payment Card Industry Data Security Standard), geschützte Gesundheitsdaten, unveröffentlichte Finanzergebnisse sowie vertrauliche Daten zu Fusionen und Übernahmen. Die am häufigsten in Unternehmensnetzwerken festgestellten Verstöße geben genau Aufschluss darüber, was Mitarbeiter übermitteln: Namen und amtliche Kennungen, Quellcode und medizinische Unterlagen.

Die wichtigsten Kontrollmaßnahmen

- Explizite „verbotene Daten"-Kategorien für Quellcode, personenbezogene, medizinische und Finanzdaten, Geheimnisse und unveröffentlichte Pläne
- Risikobewertung nach Usergruppe, Abteilung, Gerätestatus und Datensensitivität
- Dynamische Richtinienanpassung auf Basis kontextueller Signale, nicht nur statischer Zulassen/Blockieren-Regeln
- Getrennte Risikobehandlung für KI-Interaktionen mit Datei-Uploads im Vergleich zu reinen Textabfragen

>410 MIO.

VERSTÖSSE GEGEN DLP-RICHTLINIEN IM ZUSAMMENHANG MIT CHATGPT — HÄUFIGSTE ARTEN: NAMEN, SOZIALVERSICHERUNGSNUMMERN, QUELLCODE, MEDIZINISCHE DATEN—

Report von Zscaler ThreatLabz zur KI-Sicherheit 2026 (S. 6, 15)

4. Grundsatz:
SITZUNGSBASIERTE INLINE-DURCHSETZUNG VON RICHTLINIEN

Die Protokolle geben Auskunft darüber, was in der Vergangenheit passiert ist. Bis man einen Eintrag findet, der belegt, dass ein Entwickler proprietären Quellcode an ein nicht genehmigtes Modell übermittelt hat, sind die Daten bereits verloren. Die meisten Unternehmen verschärfen dieses Problem noch, indem sie DLP für Datei-Uploads in KI-Tools einsetzen, während Prompts ungeprüft bleiben. Wenn ein Entwickler Quellcode direkt in ein Prompt-Fenster eingibt, lösen die Kontrollmaßnahmen keine einzige Warnung aus. Das Wachstum der Datenübermittlungen verdeutlicht das Ausmaß dessen, was durch diese ungeschützten Kanäle fließt.

Die wichtigsten Kontrollmaßnahmen

- Inline-DLP bei Prompts, Antworten und Datei-Uploads, nicht nur auf URL/Kategorie-Ebene
- Bei der Richtliniendurchsetzung sind folgende Aktionen verfügbar: Zulassen, Warnen, Redigieren, Blockieren und Isolieren.
- Schnelle Klassifizierung führender GenAI-Anwendungen mit kontextbezogener Erkennung
- Inhaltsmoderation für themenfremde, eingeschränkte und wettbewerbssensible Interaktionen
- Browserisolierung für zulässige, aber riskante KI-Tools: Zugriff ohne vollständiges Vertrauen

93 %

DAS JÄHRLICHE WACHSTUM DES DATENÜBERTRAGUNGSVOLUMENS IM BEREICH KI IN UNTERNEHMEN ÜBERTRIFFT DAS, WAS DIE MEISTEN SICHERHEITSPROGRAMME ERFASSEN ODER KONTROLLIEREN KÖNNEN.

Report von Zscaler ThreatLabz zur KI-Sicherheit 2026 (S. 5, 14)

5. Grundsatz:
ZUGRIFF UND EXPOSITION MINIMIEREN

KI-Systeme schaffen Zugriffswege, die oft außerhalb traditioneller Identitäts- und Netzwerkmodelle liegen. Eine RAG-Pipeline (Retrieval-Augmented Generation) kann einem Modell Zugriff auf einen Dokumentenkorpus ermöglichen. Ein Agent kann Verbindungen zu internen Tools, Datensätzen, Vektorspeichern oder MCP-Servern herstellen. Diese Berechtigungen sind genauso real wie die Zugriffsrechte jedes anderen Users, auch wenn sie von den üblichen Zugriffsüberprüfungen nicht erfasst werden.

Die Minimierung des Zugriffs und der Offenlegung bedeutet, den Bereich, den KI-Systeme erreichen können, zu reduzieren und die Kommunikation zwischen KI-Komponenten einzuschränken. Agents, Datenpipelines, RAG-Komponenten und Backend-Datenquellen dürfen nur über explizit genehmigte Pfade verbunden werden. Ohne eine strikte Beschränkung und Segmentierung der Berechtigungen kann ein überberechtigter Agent oder eine kompromittierte KI-Komponente zur lateralen Ausbreitung von Bedrohungen führen.

Die wichtigsten Kontrollmaßnahmen

- Rollenbasierte Zugriffssteuerung für KI-Anwendungen, Agents, Modellkonnektoren, Datensätze, Vektorspeicher und KI-Infrastrukturkomponenten
- Regelmäßige Überprüfung und Reduzierung übermäßiger Zugriffsrechte auf vertrauliche Datensätze und KI-gestützte Systeme
- Minimale Zugriffsberechtigungen für KI-Agents: Definieren Sie, worauf sie zugreifen dürfen, und schränken Sie alles andere ein.
- Segmentierung von KI-Services, Agent-Plattformen, Datenpipelines, RAG-Komponenten und Backend-Datenquellen
- Beschränken Sie den lateralen Zugriff zwischen KI-Komponenten auf explizit definierte Pfade
- Protokollbasierte Prüfung und Protokollierung von MCP- und A2A-Traffic

Minimieren Sie das Schadenspotenzial für jeden KI-Zugriffspfad

Behandeln Sie jeden KI-Konnektor, jede Agentenberechtigung, jede RAG-Pipeline, jeden Vektorspeicher, jeden MCP-Server und jede Komponenten-zu-Komponenten-Verbindung als potenziellen Angriffspunkt. Beschränken Sie den Zugriffsbereich, segmentieren Sie die Kommunikationsbereiche, untersuchen Sie den Traffic und protokollieren Sie sämtliche Aktivitäten.

6. Grundsatz:
KONTINUIERLICHE ÜBERWACHUNG UND PRÜFBARKEIT

Die kontinuierliche Verifizierung ist das Betriebsmodell, kein erreichter Meilenstein, den man hinter sich lässt. Für KI-Systeme bedeutet dies umgehende Protokollierung und Reaktion, Echtzeitwarnungen bei Richtlinienverstößen, Nutzungs-Dashboards und einen Untersuchungsworkflow, der keine manuelle Protokollkorrelation über fünf verschiedene Tools hinweg erfordert. **Der aufsichtsrechtliche Druck hierfür ist bereits erheblich:** Das EU-Gesetz über Künstliche Intelligenz (EU AI Act), der NIST AI RMF (National Institute of Standards and Technology Artificial Intelligence Risk Management Framework)und neue Governance-Rahmenwerke für KI fordern allesamt nachweisbare, dokumentierte Kontrollen.

Die wichtigsten Kontrollmaßnahmen

- Protokollierung von Prompts und Antworten für Inline-KI-Interaktionen
- Nutzungsübersichten nach User, Abteilung und KI-App mit Risikosignal-Overlays
- Arbeitsablauf für Benachrichtigungen und Untersuchungen bei Verstößen gegen KI-Richtlinien und Anomalien
- Prüfungsreife Berichterstattung: Welche KI wird eingesetzt, welche Kontrollen werden angewendet und welche Verstöße traten auf?
- Zuordnung der Governance-Strategie zu NIST AI RMF, EU AI Act und anderen geltenden Frameworks

1	2	3	4	5
Reaktiv	**Sichtbare**	**Gesteuerte**	**Automatisiertes**	**Kontinuierliches**
Kein Inventar Ad-hoc-Blockierung	KI-Ressourcen bekannte Basisrichtlinien	rollenbasierte Richtlinien-Governance-Workflows	Coaching, kontinuierliche Einhaltung der Vorschriften	Red Teaming + Laufzeitschutzmechanismen

Programmreife →

1.–30. Tag

Transparenz schaffen

Man kann nur KI-Anwendungen sichern, die man sehen kann. In den ersten 30 Tagen geht es darum, ein transparentes Inventar und eine solide Basis an Kontrollen zu schaffen, um fundierte Entscheidungen treffen zu können, nicht darum, alles abzuriegeln.

1.–30. TAG Transparenz schaffen	31.–60. TAG Aufbau von Richtlinienreife	61.–90. TAG Automatisieren und validieren
• KI-gestützte App-Erkennung einsetzen • Rote Datenkategorien definieren • Basiszugriffsrichtlinien • Erstes Dashboard für Führungskräfte	• Rollenbasierte Zugriffsrichtlinien • Governance-Workflows live • AI-SPM-Rollout • Schutzmaßnahmen für private KI	• Automatisiertes User-Coaching • Kontinuierliche Compliance-Berichterstattung • Regelmäßiges Red Teaming • Closed-loop Guardrails

Implementierungsplan für 30/60/90 Tage im Überblick

Checkliste für die Implementierung (30 Tage)

- **Legen Sie einen regelmäßigen Überprüfungszyklus für KI-Tools fest:** neu entdeckte Tools im Traffic, eingegangene Ausnahmeanfragen und vorgenommene Richtlinienänderungen.
- Erstellen Sie das erste KI-Sicherheits-Dashboard für Führungskräfte: KI-App-Inventar (genehmigte/Schatten-Apps), DLP-Ereignisse nach Kategorie, Richtlinienmaßnahmen und Trendlinien
- Definieren Sie Notfallpläne für die häufigsten KI-bezogenen Sicherheitsvorfälle: vertrauliche Daten in Prompts, Zugriff auf blockierte Ziele, Uploads großer Dateien in KI-Tools.
- **Zentralisierung der KI-Sicherheitsprotokolle:** Useraktivität, App-Nutzung, Prompts/Antworten, Richtlinienmaßnahmen und Verstöße in einer einzigen Dashboard-Ansicht

Operationalisieren

- **Überprüfen und optimieren Sie Ihre DLP-Richtlinien anhand der Telemetriedaten der ersten zwei Wochen:** Reduzieren Sie Fehlalarme und verbessern Sie die Erkennung bestätigter Datenexpositionsmuster.
- **Aktivieren Sie die Inhaltsmoderation zur Durchsetzung der Nutzungsrichtlinien:** themenfremde Interaktionen, eingeschränkte Themen, wettbewerbsrelevante Inhalte und unangemessene Beiträge.
- **Fügen Sie eine Browserisolierung für KI-Tools in der zulässigen, aber risikobehafteten Stufe hinzu:** Benutzer können auf das Tool zugreifen, aber Uploads, Downloads und Kopieren/Einfügen sind auf Browserebene eingeschränkt
- **Erweitern Sie den DLP-Schutz auf den gesamten roten Datenbereich:** Quellcode (zusätzliche Sprachen und Muster), personenbezogene Daten (über offensichtliche Formate hinaus), PCI-Karteninhaberdaten, Gesundheitsdaten und vertrauliche Geschäftsinhalte

Datenschutz ausweiten

- **Veröffentlichen Sie eine Ampelrichtlinie für die zulässige Nutzung von KI:** Grün (erlaubt), Gelb (Nutzung mit Einschränkungen), Rot (gesperrt, mit Begründung). Machen Sie den Prozess für Ausnahmegenehmigungen transparent.
- Aktivieren Sie Inline-DLP für Prompts, beginnend mit hochzuverlässigen Erkennungsregeln für kritische Datenkategorien: Quellcodemuster, Anmeldedatenformate, PII-Schemas.
- **Beschränken Sie die risikoreichsten Aktionen auf nicht genehmigten Tools:** Uploads, große Datenblöcke in der Zwischenablage und Herunterladen von KI-Ausgaben in nicht verwaltete Speicherorte
- **Zielrichtlinien durchsetzen:** Warnung oder Sperrung nach Usergruppe und App für eindeutig nicht genehmigte oder risikoreiche KI-Ziele

Basiskontrollen anwenden

- **Erstellen Sie eine grundlegende KI-Risikoanalyse:** Welche Anwendungen greifen auf welche Datentypen zu und welche Usergruppen nutzen KI am häufigsten?
- **Definieren Sie die Kategorien roter Daten explizit:** Quellcode, personenbezogene Daten, PCI-Daten, Gesundheitsdaten (PHI), Geheimnisse und Zugangsdaten, unveröffentlichte Finanzdaten, vertrauliche Daten zu Fusionen und Übernahmen oder strategische Pläne

Schutzumfang

- Aktivieren Sie die Transparenz von Prompts für KI-Anwendungen mit dem höchsten Traffic, um zu verstehen, welche Daten über Prompts und Antworten fließen.
- Erstellen Sie eine Übersicht, welche KI-Tools in interne Systeme integriert sind und auf welche Datentypen sie zugreifen.
- **Schatten-KI erkennen:** Tools im aktiven Einsatz, die nie formell genehmigt wurden.
- Bereitstellung von KI-App-Erkennung zur Bestandsaufnahme von GenAI-SaaS, eingebetteter KI in Unternehmensplattformen und KI-Entwicklertools, die in der gesamten Umgebung verwendet werden.

ABWÄGUNG ERFORDERLICH

Der Einsatz von Warnrichtlinien anstelle von strikten Sperren reduziert Reibungsverluste und deckt Fehlalarme in Ihren DLP-Regeln auf, bevor diese legitime Arbeitsabläufe blockieren. Der Nachteil besteht darin, dass Datenlecks auch während der Warnphase noch auftreten können. Wägen Sie dies gegen Ihre Risikotoleranz für rote Daten ab; wenn das Durchsickern von Quellcode ein kritisches Risiko darstellt, sind harte Blockaden bei der Code-Mustererkennung das Problem früher Fehlalarme wert.

Ergebnisse aus den ersten 30 Tagen

- Inventar von KI-Apps mit Risikostufenklassifizierung (genehmigt, toleriert, verboten)
- Aktive Zugriffsrichtlinien gelten für KI-Anwendungen und Usergruppen mit dem höchsten Risiko
- Aktive Prompt-DLP für rote Datenkategorien nach abgeschlossener Erstkonfiguration
- Dashboard für Führungskräfte mit Einblick in KI-Inventar und Risikokennzahlen
- Leitfaden zur akzeptablen Nutzung von KI und Verfahren für Ausnahmeanträge veröffentlicht

31.–60. Tag

Aufbau von Richtlinienreife

In diesem Zeitraum wird auf der vorhandenen Transparenz und den ersten Kontrollmechanismen aufgebaut, um ein ausgereiftes Richtlinienmodell, funktionierende Governance-Workflows und ein Inventar der KI-Infrastruktur zu etablieren.

Checkliste für die Implementierung (60 Tage)

Zugriffsrichtlinienmodell weiterentwickeln

- **Abteilungsspezifische KI-Zugriffsrichtlinien:** Personalwesen, Finanzen, Entwicklung und Vertrieb haben unterschiedliche Anforderungen an KI-Tools und unterschiedliche Risikoprofile hinsichtlich der Datenoffenlegung.
- **Steuerung bedingter Richtlinienebenen:** Zugriffsebenen für verwaltete Geräte versus BYOD, geografische Zugriffssignale, Risikobewertung nach Gerätestatus
- **Formalisierung von drei Zugriffsebenen für KI-Anwendungen:** genehmigt (voller Zugriff mit Inline-Prüfung), toleriert (überwachter Zugriff mit Nutzungsbeschränkungen), verboten (blockiert mit Ausnahmemöglichkeit)
- Überprüfung und Aufhebung von Ausnahmeanträgen aus den ersten 30 Tagen im Hinblick auf das nun ausgereiftere Richtlinienmodell

Governance-Prozesse etablieren

- **KI-Governance-Anforderungen internen Richtlinien und geltenden Frameworks zuordnen:** NIST AI RMF, EU AI Act, HIPAA (sofern PHI betroffen ist).
- **Genehmigungsworkflows für KI-Tools formalisieren:** Einreichung, Sicherheitsprüfung, Risikoklassifizierung, bedingte Genehmigung und regelmäßige Revalidierung
- Legen Sie standardisierte Sensitivitätsbezeichnungen für Datentypen fest, die für die KI-Nutzung relevant sind, mit klaren Handhabungsregeln für jede Kategorie.
- **Definieren Sie Eskalationswege für KI-Sicherheitsvorfälle:** Wer ist für die Reaktion zuständig? Was gilt als Vorfall, der eine Benachrichtigung der Geschäftsleitung erfordert? Und welche Beweise werden gesichert?

Inventarisierung der KI-Infrastruktur

- Inventarisieren Sie KI-Modelle, Agents, Services, Datensätze, Vektorspeicher und Pipelines mit Herkunftsnachweis von Datensätzen bis zur Laufzeitnutzung.
- **AI-SPM-Bewertung durchführen:** Identifizieren Sie Fehlkonfigurationen, übermäßige Berechtigungen, die Offenlegung vertraulicher Daten und Schwachstellen in den Komponenten der KI-Infrastruktur.
- **Behebung von KI-SPM-Problemen mit hohem Risiko priorisieren:** übermäßige Berechtigungen, offengelegte vertrauliche Daten und unkontrollierte Zugriffspfade für KI-Komponenten.
- **Plattformübergreifenden Schutz validieren:** Cloud-basierte Modelle, unternehmensweit eingesetzte Agentenplattformen, MCP-Serverintegrationen und nicht verwaltete KI-Infrastruktur

Schutz privater KI-Anwendungen

- **Kontrollmechanismen für private KI-Anwendungen eine Erkennung von Prompt-Injection hinzufüge:** absichtlich erstellte Eingaben, die darauf abzielen, Modellanweisungen zu überschreiben oder vertrauliche Informationen zu extrahieren.
- **Implementierung eines Überwachungssystems für Data Poisoning in RAG-Pipelines:** Erkennung unautorisierter Änderungen am Dokumentenkorpus, die die Modellausgaben verfälschen könnten.
- **Protokollierung von Prompts/Antworten für private KI-Systeme mit Anomalieerkennung:** ungewöhnliche Abfragemuster, unerwartete Ausgabeinhalte und Interaktionen außerhalb des Gültigkeitsbereichs
- Etablierung von Arbeitsabläufen zur Überprüfung von Modellinteraktionen für hochsensible private KI-Anwendungen

Ergebnisse aus den ersten 60 Tagen

- Ausgereiftes rollenbasiertes Zugriffsrichtlinienmodell mit bedingter Durchsetzung (live)
- Arbeitsablauf für die KI-Governance in Betrieb: Tool-Genehmigung, Sensitivitätskennzeichnung, Eskalationswege definiert
- AI-SPM-Visualisierungsbaseline mit priorisiertem Behebungsrückstand
- AI-BOM, das die wichtigsten KI-Infrastrukturkomponenten umfasst
- Schutzmaßnahmen für interne KI-Anwendungen mit höchster Priorität
- Berichtsrhythmus für die Einhaltung der Vorschriften festgelegt und erster Bericht erstellt

61.–90. Tag

Automatisieren und validieren

In diesem Zeitraum wird das Programm selbsttragend: automatisierte Durchsetzung, kontinuierliche Überwachung der Einhaltung und Validierung durch Red Teaming.

Checkliste für die Implementierung (90 Tage)

Durchsetzung automatisieren

- **Automatisiertes User-Coaching bei Richtlinienverstößen:** kontextspezifische Warnmeldungen vor einer Blockierung.
- **ITSM- oder SOAR-Ticketing für wiederholte Verstöße:** Erstellen Sie automatisch Untersuchungstickets für Benutzer, die wiederholt dieselben Kontrollen auslösen.
- **Automatische Quarantäne- und Isolationsauslöser für risikoreiches Verhalten:** Übermittlung großer Mengen vertraulicher Daten, Zugriff auf verbotene KI-Ziele, ungewöhnliche Promptmuster
- **Feedbackschleifen aufbauen:** Trends bei Richtlinienverstößen im Zeitverlauf verfolgen, um Verhaltenskategorien zu identifizieren, die eher auf Anpassungsbedarf der Richtlinien als auf Durchsetzungsmängel hinweisen.

Sicherstellung der Compliance-Berichterstattung

- **Kontinuierliche Überwachung des KI-Compliance-Status konfigurieren:** Welche Kontrollen sind aktiv, welche Assets sind abgedeckt und wo bestehen Schutzlücken?
- **Kontrollmechanismen den anwendbaren Frameworks zuordnen:** NIST AI RMF (einschließlich NIST AI 600-1 für generative KI-Risiken), Verpflichtungen des EU-KI-Gesetzes, DSGVO und HIPAA, sofern PHI (geschützte Gesundheitsdaten) in den Anwendungsbereich fallen.
- **Erstellung von revisionssicheren Berichtspaketen:** Zusammenfassung der KI-App-Nutzung, Benachrichtigung über DLP-Maßnahmen und deren Ergebnisse, gewährte und abgelaufene Richtlinienausnahmen, Status der Behebung von KI-SPM-Ergebnissen
- **Vierteljährliche Überprüfung des KI-Sicherheitsprogramms ein:** Wirksamkeit der Richtlinien, Deckungslücken, neue Risikobereiche und Anpassungen der Roadmap.

Red Teaming und Härtung

- **Etablieren Sie einen Red-Teaming-Zyklus für intern entwickelte KI-Anwendungen:** vorgefertigte Angriffsbibliotheken plus maßgeschneiderte Testfälle, die auf Ihren spezifischen Anwendungskontext abzielen.
- **Führen Sie Tests durch, die die gesamte KI-bedingte Angriffsfläche abdecken:** Prompt-Injection, Jailbreaks, Kontextvergiftung, Extraktion personenbezogener Daten, Quellcodelecks, themenfremdes Verhalten und Regressionsanalyse von Modellbenchmarks.
- **Implementierung von Laufzeitschutzrichtlinien für produktive KI-Systeme:** Erkennung von Prompt-Injection, Datenlecks (personenbezogene Daten und Quellcode), Jailbreak-Versuchen und Verstößen gegen die Inhaltsmoderation.
- **Automatisierte Umsetzung von Red-Teaming-Ergebnissen in Laufzeit-Schutzmaßnahmen:** Die Lücke zwischen Test und Produktionsdurchsetzung ist der Punkt, an dem die meisten KI-Anwendungssicherheitsprogramme versagen.

16 Minuten	72 %	Alle
mittlere Zeit bis zum ersten kritischen Fehler bei KI-Red-Team-Tests. In über 25 Umgebungen wurden mehr als 222.000 Angriffe durchgeführt.	der Unternehmen deckten bereits beim ersten Test eine kritische Schwachstelle auf. Kritische Risiken bestehen vom ersten Tag an, selbst in ausgereiften Umgebungen.	getesteten KI-Systeme wiesen mindestens eine kritische Sicherheitslücke auf. Kein KI-System ist standardmäßig sicher. Kontinuierliche Tests sind unerlässlich.

Architekturausrichtung prüfen

- **Ordnen Sie alle KI-Zugriffspfade den sechs Zero-Trust-Prinzipien zu:** Überprüfen Sie, ob jedes Prinzip auf allen Zugriffsebenen für Mitarbeiter, Entwickler und KI-Infrastruktur angewendet wird.
- **Lücken an den Schnittstellen identifizieren und schließen:** Stellen, an denen eine Durchsetzungsebene endet und eine andere beginnt, ohne dass eine einheitliche Abdeckung gegeben ist.
- Standardisierung der Durchsetzung durch ein einheitliches Richtlinienmodell, das Identitätsprüfung, Zielortkontrolle, Risikobewertung, Inline-Durchsetzung, Zugriffs- und Gefährdungsminimierung sowie Überwachung umfasst.

Ergebnisse aus den ersten 90 Tagen

- Automatisierte Arbeitsabläufe für die Reaktion auf Verstöße und das User-Coaching
- Kontinuierliche Compliance-Überwachung in Echtzeit mit dokumentierter Rahmenausrichtung
- Governance-Berichtspaket für Führungskräfte und die interne Revision
- KI-Red-Teaming-Rhythmus etabliert, Laufzeitschutzrichtlinien aktiv
- Geschlossener Regelkreis zur Umsetzung der Ergebnisse von Red-Teaming-Tests in Produktionsrichtlinien
- Programm-Scorecard und vierteljährlicher Fahrplan für die laufende Kontrolle

Wie Zscaler die sichere Einführung von KI ermöglicht

Die Zero Trust Exchange™ ist speziell darauf ausgelegt, KI im Unternehmensmaßstab zu steuern und abzusichern. Sie wird nicht nachträglich in eine bestehende Sicherheitsarchitektur integriert, sondern erweitert die gleiche Plattform, die bereits den User-, Workload- und Netzwerkzugriff für Tausende von Unternehmen weltweit regelt.

Jede Stufe im Schwungrad ist einer bestimmten Reihe von Plattformfunktionen zugeordnet. Die folgende Tabelle zeigt, was Zscaler auf jeder Ebene liefert.

ERKENNEN Lückenlose Transparenz über das KI-Inventar	KONTROLLE Zero-Trust-Zugriff für KI erzwingen	SCHÜTZEN KI-Anwendungen härten und schützen
KI-Asset-Management Zscaler greift auf jede KI-Anwendung, jedes Modell, jeden Agent, jeden MCP-Server und jede eingebettete SaaS-KI-Funktion in Ihrer Umgebung zu, einschließlich Tools, die die IT nie genehmigt hat.	**AI Access Security** Durchsetzung von Zero-Trust-Zugriffskontrollen für KI-SaaS, eingebettete KI in Unternehmensplattformen und KI-Entwicklungsumgebungen mit Inline-Inspektion auf jeder Ebene.	**KI-Red-Teaming und KI-Guardrails** Testen Sie privat entwickelte KI-Anwendungen unter realen Angriffsbedingungen und übersetzen Sie die Ergebnisse anschließend automatisch in Laufzeitschutzrichtlinien.
• Aufdeckung und Nutzung von Schatten-KI in allen Usergruppen • **KI-Inventar:** Herkunft von Datensätzen über Modelle und Agenten bis hin zur Laufzeitnutzung • **AI-SPM:** Kontinuierliche Statusanalyse zur Erkennung von Fehlkonfigurationen, übermäßigen Berechtigungen und Offenlegung sensibler Daten • KI-Agentenerkennung für alle unternehmensweit eingesetzten und in SaaS eingebetteten Agents	• Granulare Richtlinien für Zulassung, Warnung, Isolierung und Blockierung nach Usergruppe, App und Datentyp • Inline-DLP für Prompts, Antworten und Datei-Uploads im gesamten KI-Traffic • Zero Trust Browser für unkontrollierten und BYOD-Zugriff auf KI-Tools • Schnelle Extraktion und Klassifizierung für führende GenAI-Apps	• **Automatisiertes, kontinuierliches Red Teaming:** Prompt Injection, Jailbreaks, Kontext-Poisoning, Datenlecks, Nachlässigkeit der User • Schnelle Härtung und Modell-Benchmarking mit Governance-Mapping auf NIST AI RMF, EU AI Act, OWASP LLM Top 10 • **Laufzeit-Schutzmechanismen:** Detektoren für Jailbreaks, PII/source Code-Lecks, Inhaltsmoderation und themenfremdes Verhalten • **Geschlossene Richtliniengenerierung:** Ergebnisse von Red-Teaming-Übungen werden automatisch zu Produktionsdetektoren

DER PLATTFORMVORTEIL

Die meisten Sicherheitsanbieter lösen nur einen Teil des KI-Sicherheitsproblems. Zscaler deckt den gesamten KI-Lebenszyklus auf einer einzigen Plattform ab: KI überall entdecken, mit Hilfe der Zero-Trust-Richtlinie kontrollieren, wer auf welche KI zugreift, und KI-Anwendungen und -Infrastruktur vom Build-Prozess bis zur Laufzeit schützen. Die geschlossene Regelschleife zwischen den Ergebnissen der Red-Teaming-Übungen und den Schutzmechanismen für den Produktionsbetrieb ist eine Fähigkeit, die keine Insellösung nachbilden kann.

Häufig gestellte Fragen

Was bedeutet „Zero Trust AI“ in der Praxis?
Zero Trust AI bedeutet, das Kernprinzip von Zero Trust anzuwenden — also kein implizites Vertrauen vorauszusetzen und jede KI-Interaktion in Ihrer Umgebung kontinuierlich zu überprüfen. Konkret heißt das: Keine KI-Anwendung, kein User, kein Gerät, keine Prompt und kein Konnektor wird automatisch als vertrauenswürdig eingestuft. Der Zugriff wird basierend auf verifizierter Identität, Gerätestatus, Datenvertraulichkeit und Sitzungskontext gewährt. Kontrollen werden in Echtzeit direkt angewendet. Jede Zugriffsentscheidung wird protokolliert und ist nachvollziehbar.

Was sollten wir in den ersten 30 Tagen tun, um das Risiko von Schatten-KI zu reduzieren?
Beginnen Sie mit der Analyse. KI, die man nicht einsehen kann, lässt sich nicht steuern. ThreatLabz verfolgt über 3.400 KI-Anwendungen im Unternehmenstraffic — eine Zahl, die sich innerhalb eines Jahres vervierfacht hat, vor allem aufgrund eingebetteter KI-Funktionen in bereits von der IT freigegebenen Tools. Sobald Sie Transparenz geschaffen haben, klassifizieren Sie die Anwendungen mit dem höchsten Risiko und legen Sie grundlegende Zielrichtlinien fest. Fügen Sie für Ihre risikoreichsten Datenkategorien eine sofortige Datenverlustprävention (DLP) hinzu. Veröffentlichen Sie einen transparenten Leitfaden zur zulässigen Nutzung und einen Prozess für Ausnahmegenehmigungen, damit Benutzer einen regelkonformen Weg vorfinden und nicht dazu animiert werden, Kontrollen zu umgehen.

Wie schützen KI-Zugriffskontrollen und Inline-DLP Prompts und Uploads?
Zugriffskontrollen legen fest, wer unter welchen Bedingungen von welchen Geräten und Standorten aus auf welche KI-Tools zugreifen kann. Inline-DLP arbeitet auf der Inhaltsebene und prüft die tatsächlichen Eingaben der Nutzer, Eingabeaufforderungstexte, Datei-Uploads und Kopieren/Einfügen-Interaktionen in Echtzeit. Die beiden Kontrollmechanismen ergänzen sich: Die Zugriffskontrolle beschränkt den Zugriff auf bestimmte KI-Systeme, während DLP den Datenfluss über die zulässigen Kanäle steuert. Jede Lücke in einem der beiden Systeme birgt Risiken, weshalb Unternehmen, die DLP für Uploads einsetzen, aber auf eine sofortige Überprüfung verzichten, weiterhin einem erheblichen Datenleckrisiko ausgesetzt sind.

Was ist AI-SPM und warum ist es für die KI-Einführung so wichtig?
AI-SPM ist die kontinuierliche Bewertung von KI-Infrastrukturkonfigurationen, Berechtigungen und Sicherheitsrisiken — das KI-Äquivalent zu CSPM für Cloud-Infrastrukturen. Es erkennt KI-Assets, bewertet deren Konfigurationen anhand bewährter Sicherheitspraktiken, identifiziert übermäßige Berechtigungen und die Offenlegung sensibler Daten und verfolgt die Entwicklung des Sicherheitsstatus im Zeitverlauf. Es ist wichtig, da das Potenzial für Fehlkonfigurationen in KI-Infrastrukturen groß ist und von den meisten Unternehmen, die von KI-Experimenten zur Implementierung übergehen, nur unzureichend verstanden wird.

Welche KPIs zeigen der Führungsebene am besten, dass das KI-Risiko sinkt, ohne die Produktivität zu beeinträchtigen?
Am effektivsten ist es, eine Kennzahl zur Risikominderung mit einer Kennzahl zur KI-Nutzung zu kombinieren, sodass die Führungsebene beide Trends gemeinsam betrachtet. Die sinkende Anzahl von Schatten-KI-Anwendungen bei gleichzeitig steigender Nutzung genehmigter KI-Systeme ist ein klares Zeichen dafür, dass die Governance funktioniert. Schnelle DLP-Maßnahmen (Blockieren, Warnen, Zulassen) zeigen, wie die Durchsetzung kalibriert ist. Die Bearbeitungszeit von Ausnahmeanfragen und die Genehmigungsdauer von KI-Tools belegen die Reaktionsfähigkeit der Governance.

Schneller agieren. Dauerhaft schützen.

Über Zscaler

Zscaler (NASDAQ: ZS) ist ein Pionier und weltweit führender Anbieter von Zero-Trust-Sicherheitslösungen. Die größten Unternehmen der Welt, Betreiber kritischer Infrastrukturen und Regierungsbehörden vertrauen auf Zscaler zum Schutz von Usern, Zweigstellen, Anwendungen, Daten und Geräten und zur Beschleunigung von Initiativen zur digitalen Transformation. Die Zscaler Zero Trust Exchange™, die in über 160 Rechenzentren weltweit bereitgestellt, bekämpft in Kombination mit fortschrittlicher KI täglich Milliarden von Cyberbedrohungen und Richtlinienverstößen und ermöglicht zukunftsfähigen Unternehmen Produktivitätssteigerungen durch Reduzierung von Kosten und Komplexität.

+1 408 533 0288 Zscaler, Inc. (Hauptsitz) • 120 Holger Way • San Jose, CA 95134, USA zscaler.com/de