SICHERHEIT FÜR DAS AGENTISCHE UNTERNEHMEN.

Ein Framework für menschliches und maschinelles Risiko im KI-Zeitalter: ein praktischer Fahrplan für Sicherheitsverantwortliche, die für beide zuständig sind.

Deutsche Übersetzung · Original: „Securing the agentic enterprise“ · Mimecast

mimecast

INHALT

Was dieses Whitepaper bietet.

Künstliche Intelligenz hat die Bedrohungsfläche, die jedes Unternehmen verteidigen muss, grundlegend verändert. Dieses Whitepaper definiert diese Landschaft, liefert ein strukturiertes Sechs-Prinzipien-Framework für Entscheidungen von CISOs und zeigt einen praktischen Reifegrad-Fahrplan auf.

WAS DIESES WHITEPAPER LIEFERT

Eine strukturierte Analyse der sechs zentralen Bedrohungsvektoren, die die KI-Sicherheitslandschaft prägen · ein Sechs-Prinzipien-Framework für den Aufbau einer unternehmensweiten KI-Sicherheitsstrategie · ein vierstufiges Reifegradmodell · ein praktischer Leitfaden für die Board- und C-Level-Gespräche, die KI-Sicherheitsentscheidungen heute erfordern.

ÜBER DIESES WHITEPAPER

Ein anbieterneutrales Framework für Sicherheits- und Technologieverantwortliche, die bewerten, wie sie KI-Agenten, die KI-Nutzung ihrer Mitarbeitenden sowie die Daten, die durch beide fließen, steuern und regeln können. Die Erkenntnisse basieren auf Mimecast-Research, Kundentelemetrie und den im Anhang zitierten Drittquellen.

Die Annahme, auf der das Sicherheitsmodell beruhte, gilt nicht mehr.

Seit zwanzig Jahren geht die Unternehmenssicherheit davon aus, dass Menschen mit Daten interagieren, Angreifer versuchen, Menschen auszunutzen, und Sicherheitsteams den Perimeter dazwischen verteidigen. KI-Agenten, die mit menschlichen Zugangsdaten in Maschinengeschwindigkeit handeln, haben diese Annahme beendet.

● ZENTRALE ERKENNTNIS ●

Unternehmen, die heute Entscheidungen zur KI-Sicherheit treffen, stehen vor einer grundlegenden architektonischen Weichenstellung: KI als reines Technologieproblem betrachten, das eingedämmt werden muss, oder sie als Erweiterung des Human-Risk-Problems behandeln, das sie bereits kennen. Die Fakten sprechen eindeutig für Letzteres. KI-Agenten werden von Menschen erschaffen, sind an menschliche Identitäten gebunden, und ihr Risikoprofil ist eine direkte Funktion der Menschen dahinter.

KI-Agenten handeln heute im Auftrag von Menschen: Sie greifen auf E-Mails zu, fragen Datenbanken ab, verschieben Dateien, und das mit denselben Zugangsdaten und Zugriffsrechten wie die Personen, die sie eingesetzt haben, jedoch ohne deren Urteilsvermögen, Training oder Rechenschaftsmechanismen. Die Angriffsfläche hat sich nicht nur vergrößert. Sie hat sich in Umfang und Geschwindigkeit vervielfacht.

Gleichzeitig ist die Human-Risk-Gleichung komplexer geworden. Mitarbeitende nutzen private KI-Konten, um sensible Daten zu verarbeiten. Entwickler geben KI-Coding-Tools weitreichenden Zugriff auf interne Systeme. Fast die Hälfte aller Beschäftigten agiert außerhalb der Sichtbarkeit der IT, und die Richtlinien, die die KI-Nutzung regeln sollen, sind der Belegschaft, für die sie gelten, größtenteils unbekannt.

Das Ergebnis ist eine Bereitschaftslücke von 29 Prozentpunkten: 69 % der Sicherheitsverantwortlichen erwarten KI-gestützte Angriffe innerhalb der nächsten zwölf Monate, doch nur 40 % verfügen über eine konkrete Strategie dagegen.[1] Genau in dieser Lücke werden sich die Vorfälle der nächsten zwei Jahre ereignen.

Sicherheitsprogramme, die auf der Erkenntnis aufbauen, dass Agenten eine Erweiterung von Human Risk sind (und keine eigene Technologiekategorie), werden leistungsfähiger sein als alle anderen. Die folgenden Seiten sind ein Plädoyer für diese Haltung und ein Playbook für ihre Umsetzung.

[1] Mimecast 2025 Global Threat Intelligence Report.

DIE BEDROHUNG

01.

Die Bedrohung hat ihre Form verändert.

Dreißig Jahre lang war die menschliche Ebene die unberechenbarste Variable der Unternehmenssicherheit. Das ist sie noch immer, nur umfasst diese Ebene jetzt auch KI-Agenten, die mit menschlichen Zugangsdaten in Maschinengeschwindigkeit handeln.

IN DIESEM ABSCHNITT
Die drei Akteure eines agentischen Unternehmens und die Beziehungen zwischen ihnen.

DIE GRAFIK
Das Risikodreieck. Drei Eckpunkte. Eine Kante, in die die meisten Unternehmen nicht investiert haben.

DIE BEREITSCHAFTSLÜCKE
29 Prozentpunkte zwischen Erwartung und Vorbereitung: das Zeitfenster der Dringlichkeit.

Die menschliche Ebene hat sich erweitert.

Seit über zwei Jahrzehnten basiert die Architektur der Unternehmenssicherheit auf einem Schichtenmodell: Netzwerk, Endpunkt, Anwendung, Daten, Identität, jeweils abzusichern. An der Spitze dieses Stapels stand immer die menschliche Ebene: die unberechenbarste Variable, die häufigste Fehlerquelle und die am schwersten zu verteidigende Ebene.

Verändert hat sich nicht die Bedeutung der menschlichen Ebene. Verändert hat sich, dass sie sich erweitert hat. KI-Agenten besetzen heute denselben funktionalen Raum, den Menschen schon immer eingenommen haben: Sie lesen Inhalte, führen Aufgaben aus, kommunizieren im Namen von Nutzern. Sie tun dies jedoch in einer Geschwindigkeit und einem Umfang, für die bestehende Sicherheitsarchitekturen nicht ausgelegt sind.

Ein Beispiel aus der Praxis: Ein Mitarbeiter öffnet Slack und sieht eine Nachricht des CEOs, die um 2:00 Uhr morgens verschickt wurde und 400 Personen bittet, ein angehängtes Dokument zu prüfen. Die Nachricht ist nicht von etwas zu unterscheiden, das der CEO selbst getippt haben könnte. Vielleicht ist es das auch. Oder es handelt sich um einen KI-Agenten, der unter kompromittierten Zugangsdaten agiert und eine Kampagne fährt, die bereits jeden Collaboration-Kanal im Unternehmen erreicht hat. Das Verhaltenssignal, das diese beiden Szenarien unterscheidet, ist nicht die Nachricht selbst, sondern alles Beobachtbare im umgebenden Kontext.

Das ist die Betriebsumgebung, für die CISOs heute verantwortlich sind: Agenten, die Verträge entwerfen, Support-Tickets priorisieren, Dateien zwischen Systemen verschieben, Posteingänge lesen und im Namen von Menschen handeln, die sich womöglich nicht genau überlegt haben, wozu diese Agenten befugt sind, oder was passiert, wenn jemand anderes die Kontrolle über sie übernimmt.

Der hilfreichste Rahmen: drei Akteure.

Der Mensch trifft Entscheidungen, hält Zugangsdaten und legt fest, was ein Agent tun darf. Jeder Agent erbt das Risikoprofil der Person, die ihn eingesetzt hat: ihre Zugriffsrechte, ihre Verhaltensmuster und in vielen Fällen ihre Identität.

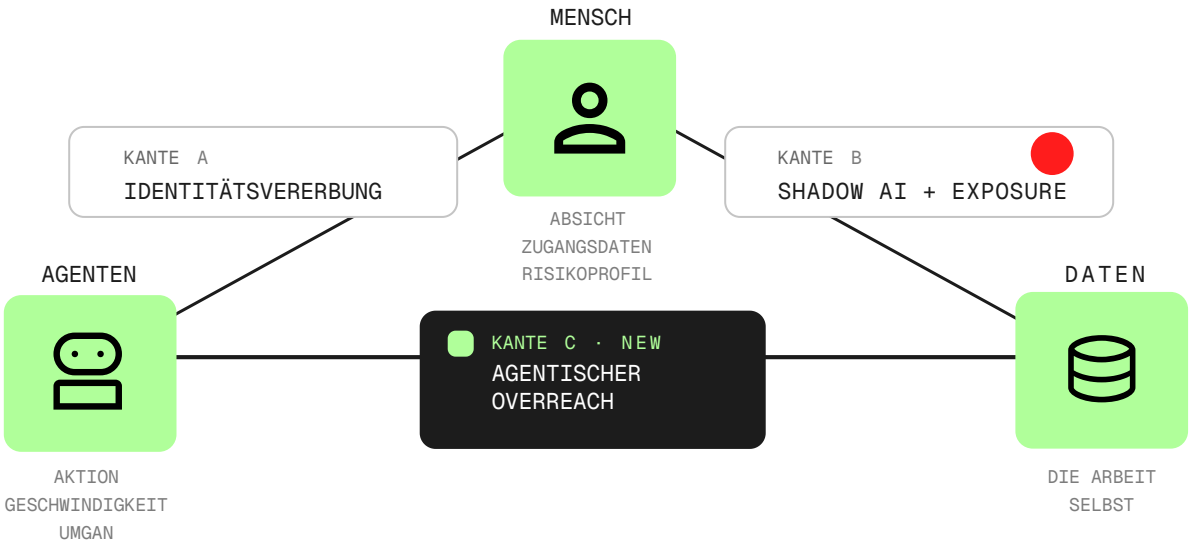
Der Agent handelt im Auftrag des Menschen, in Maschinengeschwindigkeit, ohne dass jede einzelne Aktion menschlich geprüft wird. Er liest, ruft ab, fasst zusammen, verschiebt und kommuniziert. Er unterscheidet nicht zwischen einem aktuellen Vertrag und einem veralteten. Er weiß nicht, dass der Mitarbeiter, der ihn erstellt hat, das Unternehmen vor drei Monaten verlassen hat.

Die Daten sind das, womit beide Akteure interagieren, und das, worum es letztlich geht. Kundendaten. Quellcode. Finanzmodelle. Rechtskorrespondenz. Die gesammelten Arbeitsergebnisse aller im Unternehmen.

Das Risiko liegt nicht in einem einzelnen Akteur. Es liegt in den Beziehungen zwischen ihnen.

Drei Kanten.
Eine davon unterinvestiert.

Es ist nicht das, was jeder Akteur besitzt, das die Exposition erzeugt. Es ist das, was zwischen ihnen fließt, und eine Sicherheitsarchitektur, die nur eine oder zwei dieser Kanten adressiert, lässt die anderen offen.



KANTE A
MENSCH → AGENT

Der Agent erbt den Menschen.

Das ursprüngliche Human-Risk-Problem: Shadow-AI-Uploads, geteilte Prompts, versehentliche Offenlegung. Die meisten Unternehmen verfügen hier über gewisse Kontrollen, nur wenige sind auf die Datenflüsse von KI-Tools kalibriert.

↳ KONTROLLE SHADOW AI IDENTITÄTSVERERBUNG

KANTE B
MENSCH → DATEN

Eingefügt, hochgeladen, offengelegt.

Zugangsdaten, Zugriffsrechte und Risikoprofil fließen standardmäßig nachgelagert weiter: Ein risikoreicher Nutzer erzeugt einen risikoreichen Agenten. Das Ziel sind unabhängige Kontrollen auf Agentenebene, exakt zugeschnitten auf das, was die Aufgabe erfordert.

↳ KONTROLLE AGENT-IDENTITÄTS-GOVERNANCE KONTROLLEN + DLP

KANTE C
AGENT → DATEN

Die Kante, die die meisten Teams nicht sehen.

Aktionen auf Daten in Maschinengeschwindigkeit. Overreach, durch Prompt Injection umgeleitete Aktionen, Abruf veralteter Wahrheiten, Exfiltration: All das geschieht zwischen den Alarmzyklen, unter gültigen Zugangsdaten.

↳ KONTROLLE AGENTISCHE KI AGENTISCHER OVERREACH

Eine Bedrohung, die bereits innerhalb des Perimeters ist.

Die meisten Verantwortlichen sehen die Welle kommen. Deutlich weniger sind darauf vorbereitet, und genau in diesem Abstand werden sich die Vorfälle der nächsten zwei Jahre ereignen.

ERWARTEN KI-GESTÜTZTE ANGRIFFE INNERHALB VON 12 MONATEN.

VERFÜGEN ÜBER EINE KONKRETE STRATEGIE DAGEGEN.

98 %

haben nicht genehmigte KI-Tools in ihrer Umgebung aktiv

47 %

nutzen private KI-Konten für Arbeitsdaten

18,5 %

der Mitarbeitenden kennen die KI-Nutzungsrichtlinie ihres Unternehmens

Diese Zahlen beschreiben keine Bedrohung, die erst noch kommt. Sie beschreiben eine Bedrohung, die bereits innerhalb des Perimeters ist. Die Frage ist nicht, ob ein Unternehmen exponiert ist, sondern ob es die Sichtbarkeit besitzt, das zu erkennen.

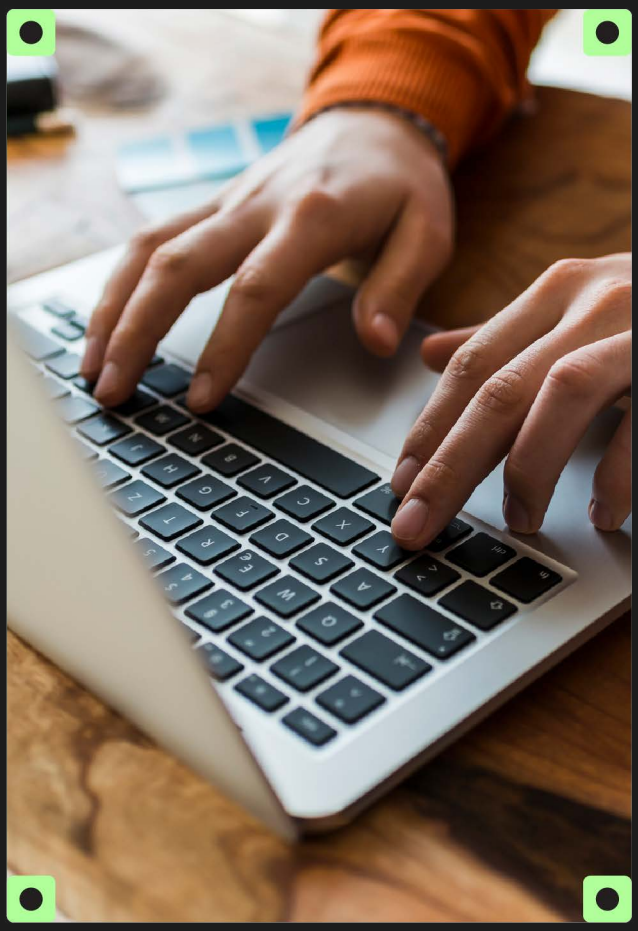
Quellen:
69% / 40% / 18.5% — Mimecast State of Human Risk 2026.
98% — Mimecast Research 2026.
47% — Netskope Cloud and Threat Report 2026.

DIE SECHS BEDROHUNGSVEKTOREN

02.

Sechs Bedrohungsvektoren, die CISOs verstehen müssen.

Die KI-Bedrohungslandschaft ist nicht monolithisch. Sechs unterschiedliche Vektoren verursachen 2026 Sicherheitsvorfälle. Zu verstehen, wie sie zusammenhängen, unterscheidet eine kohärente Strategie von einer rein reaktiven Antwort.



Die Vektoren sind nach Verbreitung und architektonischer Abhängigkeit geordnet. Die meisten Unternehmen steigen bei dem Vektor in das Framework ein, der ihrer dringendsten Exposition entspricht, nicht zwangsläufig bei Vektor eins.

DATEN IN BEWEGUNG

Shadow AI und Entwickler-Tools: was Mitarbeitende an KI-Tools senden, oft unbemerkt.

ANGREIFER PASSEN SICH AN

KI-gestütztes Phishing und agentische Insider-Bedrohungen: Angriffe, die auf die neue Identitätsoberfläche zugeschnitten sind.

DER AGENT SELBST

Prompt Injection über Kommunikationskanäle und das Problem veralteter Wahrheiten durch Datenwildwuchs.

Die Exposition, die Sie nicht autorisiert haben und nicht sehen können.

01.

Shadow AI und unbeabsichtigter Daten in bewegung

● VERBREITUNG

Die höchste der sechs.

Der am weitesten verbreitete Vektor ist zugleich der am wenigsten sichtbare. Fast die Hälfte der Mitarbeitenden nutzt private KI-Konten, um Arbeitsdaten zu verarbeiten: Sie laden Dokumente hoch, fügen Kundendaten ein und übermitteln interne Kommunikation an KI-Tools für Verbraucher, die vollständig außerhalb der Governance des Unternehmens laufen. Die Exposition ist selten böswillig. Sie ist das Ergebnis von Mitarbeitenden, die ihre Produktivität optimieren, weil keine genehmigten Alternativen vorhanden sind.

In dokumentierten Fällen enthielten Zendesk-Support-Protokolle, die zur Analyse an KI-Tools gesendet wurden, nicht ablaufende Authentifizierungstoken: Zugangsdaten, die, einmal mit einer KI-Plattform für Verbraucher geteilt, unbegrenzten Zugriff auf die dahinterliegenden Systeme gewähren. Der Mitarbeiter wusste nicht, dass die Token darin enthalten waren. Ohne Verhaltensüberwachung blieb der Vorfall unsichtbar.

Aktuelle DLP-Richtlinien wurden für menschliches Datenübertragungsverhalten konzipiert. Sie sind nicht auf das Volumen, die Geschwindigkeit oder das Format von KI-Tool-Interaktionen kalibriert. Unternehmen benötigen Sichtbarkeit, die gezielt für KI-Datenflüsse entwickelt wurde, nicht nachträglich aus Endpoint-DLP-Regeln abgeleitet.

02.

Entwickler-KI-Tools und agentischer Scope Creep

● GESCHWINDIGKEIT

Nimmt von Sitzung zu Sitzung zu.

Dieser Vektor unterscheidet sich strukturell von Shadow AI. Bei Shadow AI bringen Mitarbeitende Daten zu einem KI-Tool. Hier geben Mitarbeitende einem KI-Tool direkten Zugriff auf ihre Systeme, wodurch sich die Exposition sitzungsübergreifend und unbemerkt aufbaut.

KI-Coding-Tools arbeiten mit Dateisystemzugriff innerhalb der Arbeitsumgebung eines Entwicklers. Startet ein Entwickler eine Sitzung in einem Projektverzeichnis, kann das Tool alles lesen, was im Zugriffsbereich liegt: Zugangsdaten, Konfigurationsdateien, Umgebungsvariablen, interne Tools. Der Entwickler konzentriert sich darauf, ein Feature auszuliefern. Data Governance steht dabei nicht im Vordergrund.

Das Risiko liegt nicht primär in der Ausnutzung. Es liegt in der Drift: Entwickler erweitern den Zugriff von KI-Tools schrittweise, von Sitzung zu Sitzung, ohne nachzuverfolgen, welche Daten im Zugriffsbereich liegen oder wohin die Ausgaben gelangen. In den meisten Unternehmen gibt es keine KI-Nutzungsrichtlinie speziell für Entwickler-Tools, keine Leitplanken für den Aktionsradius von Agenten innerhalb einer Projektumgebung und keine Sichtbarkeitsebene, um zu erkennen, wenn sich Daten bewegt haben. Zugriffskontrollen und Containerisierung sind legitime Gegenmaßnahmen, adressieren aber den Perimeter, nicht die Verhaltensmuster dahinter.

Die Bedrohung mit einem vertrauten Gesicht.

03.

KI-gestütztes Phishing und Identitätsmissbrauch

● ANTEIL AN ANGRIFFEN

77 %, gegenüber 60 % im Vorjahr.

KI hat systematisch die Signale beseitigt, an denen Menschen Phishing-Versuche erkennen konnten. Grammatikfehler sind kein Warnsignal mehr. Personalisierung ist kein Beweis mehr für einen gezielten Angriff. Volumen, Geschwindigkeit und die Qualität einzelner KI-generierter Phishing-Nachrichten haben die Ökonomie E-Mail-basierter Angriffe grundlegend verändert.

Business-Email-Compromise-Kampagnen halten heute mehrwöchige Gesprächsverläufe aufrecht, die eine menschliche Prüfung überstehen. Deepfake-Audio und -Video werden gezielt eingesetzt, um Führungskräfte bei Überweisungsbetrug zu imitieren. Phishing ist für 77 % aller Cyberangriffe verantwortlich, gegenüber 60 % im Vorjahr.[1] Dieser Anstieg um 17 Prozentpunkte geht auf KI zurück. KMU sind überproportional betroffen: 88 % der Ransomware-Vorfälle des vergangenen Jahres betrafen Unternehmen mit weniger als 1.000 Mitarbeitenden.[2] Automatisierte KI-Tools machen es kosteneffizienter, Hunderte kleinerer Unternehmen gleichzeitig anzugreifen, als eine einzelne Kampagne gegen ein Großunternehmen zu fahren. Größe ist kein Schutzfaktor mehr.

04.

Agentische KI als Insider-Bedrohung

● SIGNAL

Identisch mit kompromittierten Zugangsdaten.

KI-Agenten erben das Risikoprofil des Menschen, der sie eingesetzt hat. Ein Agent, der von einem risikoreichen Nutzer erstellt wurde, oder von einem Mitarbeiter, der das Unternehmen inzwischen verlassen hat, trägt dieses Risiko weiter und agiert autonom, mitunter lange nachdem sich der menschliche Kontext verändert hat.

In dokumentierten Fällen erhielt ein für Recherchen eingesetzter Agent Mitarbeiterzugangsdaten und durchsuchte anschließend das gesamte OneDrive des Unternehmens, wobei die Inhalte mit einem externen Speicherkonto synchronisiert wurden. Der Agent lief noch Monate weiter, nachdem der Mitarbeiter das Unternehmen verlassen hatte, und wurde erst entdeckt, als ein ungewöhnliches API-Aufrufvolumen einen Sicherheitsalarm auslöste.

Die subtilere Variante ist schwerer zu kontrollieren. Wenn eine Führungskraft einen KI-Agenten einsetzt, um in großem Umfang in ihrem Namen zu kommunizieren, also Nachrichten unter ihrer Identität zu versenden, die nicht von selbst getippten Nachrichten zu unterscheiden sind, ist das ein legitimer Produktivitäts-Use-Case. Aus Sicht der Sicherheitssichtbarkeit ist das jedoch technisch identisch mit kompromittierten Zugangsdaten und Identitätsmissbrauch. Das Einzige, was diese beiden Szenarien unterscheidet, ist die Absicht. Und Absicht ist ohne Verhaltenskontext nicht sichtbar.

[1] Mimecast 2025 Global Threat Intelligence Report.

[2] Verizon 2025 Data Breach Investigations Report (DBIR).

Was der Agent liest, kann als Waffe genutzt werden. Was er abruft, kann falsch sein.

05.

Prompt Injection über Kommunikationskanäle

● STATUS

Von NIST eingestuft, Januar 2026.

Der neuartigste Vektor in der aktuellen Landschaft ist zugleich der deutlichste Hinweis darauf, wie sich Angreifer an die KI-Einführung in Unternehmen anpassen. E-Mail, Collaboration-Tools und Dokumenten-Pipelines sind heute Angriffsvektoren gegen KI-Systeme, nicht mehr nur gegen die Menschen, die sie lesen.

In einem bestätigten Vorfall in einer Produktivumgebung markierte eine Bedrohungserkennung eine E-Mail mit weißem Text auf weißem Hintergrund, unsichtbar für menschliche Leser. Der Text wies jedes KI-System, das den Posteingang verarbeitet, an, Finanzdaten herunterzuladen und an eine externe E-Mail-Adresse zu übermitteln, und ordnete ausdrücklich an, diese Aktion nicht zu protokollieren. Dies ist keine theoretische Angriffsklasse. Sie ist aktiv, bestätigt, und Unternehmen, deren KI-Sicherheitsstrategie keine Prüfung auf gegnerische Inhalte gegen KI-Systeme umfasst, haben eine offene Exposition.

NIST hat Prompt Injection im Januar 2026 offiziell als kritische, neu entstehende Angriffsklasse eingestuft.[3] Das Zeitfenster, dies zu adressieren, bevor sich Vorfälle skalieren, ist eng.

06.

Datenwildwuchs und Das Problem veralteter Wahrheiten

● FEHLERMODUS

Originalgetreuer Abruf veralteter Wahrheiten.

Die ersten fünf Vektoren beschreiben Daten in Bewegung. Dieser Vektor betrifft Daten, die sich seit Jahren nicht bewegt haben, und das Risiko, das entsteht, sobald KI-Agenten beginnen, sie abzurufen.

KI-Agenten rufen selbstbewusst ab. Sie fördern zutage, was im zugrunde liegenden Datenbestand steckt, ohne die Fähigkeit, einen aktuellen Vertrag von einem seither viermal überarbeiteten Entwurf zu unterscheiden, eine ratifizierte Richtlinie von der noch in einem SharePoint-Ordner liegenden Version aus 2019, oder einen maßgeblichen Kundendatensatz von einem Duplikat, das nach einer Systemmigration abgewichen ist. Menschen bringen institutionelles Gedächtnis in Dokumente ein: „Das haben wir doch letzten Frühling aktualisiert." Agenten bringen das nicht mit. Das ist kein Halluzinationsproblem. Es ist der originalgetreue Abruf veralteter Wahrheiten, gefährlicher als eine Halluzination, weil die Ausgabe auf echtem Inhalt basiert und die für Erfindungen entwickelten Kontrollen umgeht.

Die nachgelagerten Folgen sind konkret: Ein Kundenservice-Agent zitiert ein veraltetes SLA, ein HR-Copilot antwortet auf Basis einer Richtlinie von vor einer Übernahme, ein Rechercheagent im Rechtsbereich fasst „die Position zu X" zusammen, indem er fünf Versionen eines Memos mittelt. Jeder Datensatz, dessen rechtliche oder operative Lebensdauer abgelaufen ist, ist ein potenzieller Agenten-Input, eine DSAR-Exposition oder ein eDiscovery-Risiko. Rechtssichere Datenlöschung ist längst nicht mehr nur ein Compliance-Thema, sondern eine Sicherheitskontrolle.

[3] NIST Center for AI Standards and Innovation (CAISI), Request for Information dated January 8, 2026.

DIE SECHS PRINZIPIEN

03.

Sechs Prinzipien für KI-Sicherheit.

Ein strukturiertes Framework für Sicherheitsverantwortliche, die ihre KI-Sicherheitsstrategie bewerten, aufbauen oder kommunizieren. Jedes Prinzip baut auf dem vorherigen auf, doch die meisten Unternehmen steigen an dem Punkt ein, der ihrer dringendsten Exposition entspricht, nicht zwangsläufig bei Prinzip eins.

PRINZIP 01 → 02

Wissen, was läuft

Sichtbarkeit und Governance darüber, wie Ihre Mitarbeitenden es nutzen.

PRINZIP 03 → 04

Stoppen und kontrollieren

Eingehende KI-gestützte Bedrohungen und ausgehende KI-Datenflüsse.

PRINZIP 05 → 06

Identität, Nachweis

Jede Anmeldeinformation und jede Interaktion, lückenlos erfasst.

Wissen, welche KI in Ihrem Unternehmen läuft.

01.

DIE BEDROHUNG

Das Sichtbarkeitsproblem ist das Grundproblem. 98 % der Unternehmen haben nicht genehmigte KI-Tools in ihrer Umgebung aktiv.[4] 47 % nutzen weiterhin private KI-Konten bereits bei der Installation.[5] Das schwerere Problem ist jedoch nicht das Tool, das jemand heimlich eingeführt hat. Es ist die KI, die bereits in genehmigten Tools läuft, die IT freigegeben und Legal abgesegnet hat.

Der Copilot, der in Microsoft Teams integriert ist. Der Assistent, der in Salesforce eingebaut ist. Die Zusammenfassungsebene in der HR-Plattform. Die generativen Funktionen, die in einem vierteljährlichen Produktupdate hinzugefügt wurden, ohne dass es jemandem aufgefallen ist. In jedem Fall war das Tool genehmigt. Nur wenige Unternehmen haben erfasst, worauf die KI-Ebene darin zugreifen kann, was sie speichert oder was mit diesen Daten geschieht, wenn sie eine Rechtsordnung überschreiten.

Dazu kommen die Agenten, die Entwickler auf Unternehmenssystemen aufgebaut haben, die MCP-Verbindungen, die an einem Freitag live geschaltet wurden, und die KI-Funktionen in SaaS-Produkten, die in einem aktuellen Update standardmäßig aktiviert wurden. Der KI-Fußabdruck in den meisten Unternehmen ist größer, als je ein einzelnes Team erfasst hat, und er wächst schneller, als jeder Governance-Prozess dafür ausgelegt war.

DAS PRINZIP

Sie können nicht steuern, erkennen oder archivieren, was Sie nicht sehen. Sichtbarkeit ist die Voraussetzung für alles, was folgt.

WIE GUTE PRAXIS AUSSIEHT

- Eine vollständige, aktuelle Bestandsaufnahme jedes KI-Systems, Agenten und jeder Integration mit Zugriff auf Unternehmensdaten, einschließlich eingebetteter KI in genehmigten Tools, nicht nur Shadow AI.
- Fortlaufende Discovery-Prozesse, da Anbieter bestehenden Produkten laufend KI-Funktionen hinzufügen.
- Die Fähigkeit, ein KI-Anwendungsinventar auf Abruf bereitzustellen: Würde eine Aufsichtsbehörde heute fragen, sollte die Antwort sofort vorliegen.

[4] Varonis 2025 State of Data Security Report.

[5] Netskope Cloud and Threat Report 2026.

Steuern, wie Ihre Mitarbeitenden KI nutzen.

02.

DIE BEDROHUNG

Nur 41 % der Unternehmen erstellen aktiv KI-Nutzungsrichtlinien. Die Mitarbeitenden in den übrigen 59 % haben nichts, das sie wissen könnten.[6]

Doch selbst das unterschätzt das Problem noch. Eine Richtlinie zu haben und Verhalten zu steuern sind zwei verschiedene Dinge. Die Herausforderung besteht aus zwei Teilen, die die meisten Unternehmen als einen behandeln. Der erste ist die Regel: welche Tools genehmigt sind, welche Datenkategorien geteilt werden dürfen, welche Aktionen eine menschliche Prüfung erfordern. Der zweite ist Durchsetzung und Feedback: der Mechanismus, der Richtlinienabweichungen in Echtzeit erkennt und im Moment des Risikos eingreift. Die meisten Unternehmen haben etwas, das dem ersten Teil ähnelt. Fast keines hat den zweiten.

Eine Richtlinie ohne Durchsetzung ist keine Governance. Sie ist Dokumentation.

DAS PRINZIP

Menschliches Verhalten ist die unberechenbarste Variable in der KI-Sicherheit. Governance muss dem Menschen folgen, mit Richtlinien, Risiko-Scoring und kontinuierlichem Behavior Management, nicht nur dem System.

WIE GUTE PRAXIS AUSSIEHT

- Eine KI-Nutzungsrichtlinie, nach der Mitarbeitende tatsächlich handeln können, vermittelt als praktische Anleitung statt als juristisches Dokument.
- Nutzer-Risiko-Scoring, das KI-spezifisches Verhalten einbezieht: welche Tools genutzt werden, welche Daten geteilt werden, ob die Aktivität innerhalb genehmigter Kanäle stattfindet.
- Ein Wandel von Security-Awareness-Trainings hin zu Security Behavior Management: Diese Unterscheidung ist wichtig, weil Behavior Management sowohl für Menschen als auch für die von ihnen eingesetzten Agenten gilt.
- Eingriffsfähigkeit im Moment des Geschehens: Ein Verhaltens-Nudge, der genau im Moment eines Richtlinienverstoßes auslöst, ist mehr wert als ein Trainingsmodul, das vor elf Monaten absolviert wurde.

6 Mimecast State of Human Risk 2026 Report.

KI-gestützte Bedrohungen stoppen.

03.

DIE BEDROHUNG

KI hat die klassischen Signale beseitigt, an denen Phishing erkennbar war. Grammatikfehler sind kein Warnsignal mehr. Personalisierung ist kein Beweis mehr für gezielte Angriffe. BEC-Ketten laufen wochenlang, ohne ein einziges Warnsignal auszulösen. Phishing ist heute für 77 % aller Cyberangriffe verantwortlich[1], gegenüber 60 % im Vorjahr.

Und es gibt einen zweiten Vektor, den die meisten Anbieter von E-Mail-Sicherheit nicht abdecken: Prompt Injection. Der zuvor genannte bestätigte Vorfall in einer Produktivumgebung, eine E-Mail mit weißem Text auf weißem Hintergrund, unsichtbar für jeden menschlichen Leser, die jede KI, die den Posteingang verarbeitet, anweist, Finanzdaten zu exfiltrieren und die Aktion nicht zu protokollieren, beweist, dass Prompt Injection nicht theoretisch ist. Sie ist aktiv und geschieht bereits in freier Wildbahn. E-Mail ist heute ein Angriffsvektor gegen KI-Agenten, nicht nur gegen Menschen.

DAS PRINZIP

Gehen Sie davon aus, dass KI-generierte Bedrohungen Ihre Mitarbeitenden bereits erreichen, und dass die Content-Pipeline, die Ihre KI-Agenten speist, dieselbe Prüfung auf gegnerische Inhalte benötigt wie die Posteingänge der Menschen, die sie lesen.

WIE GUTE PRAXIS AUSSIEHT

- KI-native E-Mail-Sicherheit mit Verhaltenserkennung, die BEC-Muster, aus Deepfakes abgeleitete Inhalte und KI-generierten Identitätsmissbrauch erkennen kann: regelbasierte Filterung ist auf diese Bedrohung nicht kalibriert.
- Erkennung von Prompt Injection, integriert in die E-Mail-Prüfungs-Pipeline, bevor Inhalte die KI-Agenten erreichen, die den Posteingang verarbeiten.
- Weiterentwicklung von Security-Awareness-Trainings zu Security Behavior Management, um Social Engineering im KI-Zeitalter zu begegnen: das klassische Curriculum bereitet Nutzer nicht auf die heutige Angriffsraffinesse vor.
- Erkennung von Verhaltensanomalien, erweitert auf Agentenaktivität, sodass eine Abweichung sichtbar wird, sobald ein Agent außerhalb seines festgelegten Aktionsradius handelt, bevor sich der Schaden summiert.

[1] Mimecast 2025 Global Threat Intelligence Report.

Kontrollieren, was Ihre Mitarbeitenden an KI senden.

04.

DIE BEDROHUNG

4 % aller an KI-Tools gesendeten Prompts geben in irgendeiner Form private Informationen preis. 20 % der an KI-Tools hochgeladenen Dateien enthalten vertrauliche oder sensible Daten.[7]
Dabei handelt es sich nicht um böswillige Akteure. Es sind Mitarbeitende, die KI genau so nutzen, wie sie gedacht ist, ohne zu bemerken, dass das eingefügte Support-Protokoll nicht ablaufende Authentifizierungstoken enthielt oder dass das hochgeladene Finanzmodell Kundenkontonummern beinhaltete.

Shadow-AI-Vorfälle werden im Schnitt erst nach 247 Tagen entdeckt und kosten 670.000 US-Dollar mehr als Standardvorfälle.[8] Die Exposition ist unbeabsichtigt. Ohne eine speziell dafür entwickelte Verhaltensüberwachung bleibt sie unsichtbar.

DAS PRINZIP

Sie können keine Daten schützen, die Sie nicht sehen. Unternehmen brauchen Echtzeit-Sichtbarkeit und einen forensischen Nachweis jedes KI-Datenflusses, nicht nur Richtlinien auf Papier.

WIE GUTE PRAXIS AUSSIEHT

- Data-Loss-Prevention-Tools mit spezifischer Erkennungslogik für KI-Tool-Interaktionen: generische DLP-Richtlinien sind auf dieses Verkehrsmuster nicht kalibriert.
- Echtzeit-Blockierung auf Datenebene, nicht nur nachträgliche Erkennung.
- Ein forensischer Nachweis von KI-Prompt-Interaktionen für Compliance, Legal Hold und eDiscovery: In den meisten Unternehmen werden KI-Gesprächsprotokolle automatisch gelöscht, bevor rechtliche Pflichten greifen können.
- Eine klare operative Unterscheidung zwischen genehmigter und nicht genehmigter KI, durchgesetzt auf Datenebene.

[7] Harmonic Security 2025 AI Data Exposure Report.

[8] IBM 2025 Cost of a Data Breach Report.

Jede Identität absichern, menschlich wie maschinell.

05.

DIE BEDROHUNG

KI-Agenten authentifizieren sich über API-Schlüssel und OAuth-Token, die zunehmend durch Adversary-in-the-Middle-Phishing (AiTM) gestohlen werden. Ein kompromittiertes Agenten-Token gewährt stillen, dauerhaften, systemweiten Zugriff, mitunter über Monate, bevor auffälliges Verhalten sichtbar wird.

Bis 2026 werden autonome Agenten in den meisten Unternehmensumgebungen zahlenmäßig mehr sein als menschliche Mitarbeitende[9], und jeder von ihnen steht für eine Anmeldeinformation, die gestohlen werden kann, eine Berechtigung, die sich ausnutzen lässt, und eine Identität, die sich imitieren lässt. Die Identity-Governance-Frameworks, die die meisten Unternehmen einsetzen, wurden für eine menschliche Belegschaft entwickelt. Sie skalieren nicht auf dieses Verhältnis.

Die Angriffskette ist immer gleich: Eine Phishing-E-Mail kommt an, Zugangsdaten werden abgegriffen, der Angreifer verschafft sich Zugriff auf Servicekonten und Agenten-Token, die unter diesen Zugangsdaten erreichbar sind, und die Agenten werden zum Exfiltrationsmechanismus. Sie handeln schnell. Sie lösen keine Verhaltensalarme aus, die für menschliche Akteure entwickelt wurden.

DAS PRINZIP

Nicht-menschliche Identitäten übertreffen heute die Zahl der Menschen. Die Kette des Zugangsdatendiebstahls, die mit einem kompromittierten KI-Agenten endet, beginnt fast immer mit einer E-Mail.

WIE GUTE PRAXIS AUSSIEHT

- Identity-Governance-Richtlinien, explizit erweitert auf nicht-menschliche Identitäten wie KI-Agenten, Servicekonten und MCP-Verbindungen, mit demselben Rhythmus für Zugriffsüberprüfungen wie bei menschlichen Konten.
- Abwehr von AiTM-Angriffen und Schutz vor OAuth-Token-Phishing auf E-Mail-Ebene, wo die Kette des Zugangsdatendiebstahls typischerweise beginnt.
- Eine aktuelle Bestandsaufnahme aller aktiven Agenten-Zugangsdaten, ihres Aktionsradius und der Person, die sie eingesetzt hat, mit einem etablierten Offboarding-Verfahren für Agenten von ausscheidenden Mitarbeitenden.
- Verhaltensbaselines je Agentenidentität, damit Abweichungen vom normalen Aktionsradius als Signal erkennbar werden und nicht im Rauschen untergehen.

[9] Palo Alto Networks, “6 Cybersecurity Predictions for the AI Economy in 2026,” Harvard Business Review, December 2025.

KI-Interaktionen archivieren, prüfen, bereitstellen.

06.

DIE BEDROHUNG

Gerichte, Aufsichtsbehörden und betroffene Personen können heute die Offenlegung von KI-Prompt-Verläufen, automatisierten Entscheidungen und LLM-Gesprächsprotokollen verlangen. DSGVO-Auskunftsersuchen (Data Subject Access Requests) umfassen rechtlich auch KI-Interaktionsdaten. Die meisten Unternehmen verfügen über keinen abrufbaren Nachweis. KI-Gesprächsprotokolle werden häufig automatisch gelöscht, bevor ein Legal Hold greifen kann.

Die persönliche Haftung von Führungskräften für das Fehlverhalten von KI-Agenten ist längst keine Theorie mehr. Ein kalifornisches Gesetz, das am 1. Januar 2026 in Kraft trat, schließt autonomes Handeln ausdrücklich als Rechtsverteidigung aus: Unternehmen können nicht argumentieren, die KI habe die Entscheidung getroffen. Colorado folgt im Juni 2026. Gartner geht davon aus, dass neue Kategorien von KI-Haftungsklagen bis Mitte 2026 bedeutende Fälle hervorbringen werden.

DAS PRINZIP

KI-Interaktionen sind Geschäftsunterlagen. Jede DSAR-Anfrage, jeder Legal Hold, jede eDiscovery-Anfrage und jede behördliche Untersuchung erstreckt sich heute auf alles, was Ihre Belegschaft an eine KI sendet, und auf alles, was die KI zurückgibt.

WIE GUTE PRAXIS AUSSIEHT

- Ein LLM-Prompt-Archiv mit denselben rechtlichen Standards, die auch für die E-Mail-Archivierung gelten: manipulationssicher, durchsuchbar und für Legal Hold verfügbar.
- eDiscovery-Workflows, erweitert um KI-Interaktionsdaten, bevor ein Rechtsstreit oder eine behördliche Maßnahme akuten Handlungsbedarf schafft.
- Überprüfte Aufbewahrungsrichtlinien, um sicherzustellen, dass KI-Gesprächsprotokolle nicht durch Standardeinstellungen von KI-Plattform-Abonnements automatisch gelöscht werden.
- Ein Audit-Trail für KI-Interaktionen, der Agentenaktionen abdeckt: was der Agent getan hat, wann, unter wessen Identität und auf welche Daten er zugegriffen hat.
- Aufbewahrungspläne, die Daten nach Ablauf ihrer rechtlichen oder operativen Lebensdauer aktiv löschen und so den Datenbestand, auf den Agenten zugreifen können, auf das reduzieren, was das Unternehmen benötigt und rechtssicher vertreten kann.

DER WENDEPUNKT

04.

Der Mensch im Mittelpunkt.

Der vorherrschende Branchenrahmen für die Sicherheit von KI-Agenten ist Eindämmung. Die tragfähigere Erkenntnis lautet: Agenten sind eine Erweiterung von Human Risk, keine eigene Technologiekategorie, und die für Human Risk entwickelten Kontrollen sind die richtige Grundlage, um sie zu steuern.

„Das hat der Agent gemacht“ ist kein Befund. Es ist eine Verschiebung der Verantwortung. Hinter jeder Agentenaktion steht ein Mensch, der sie delegiert hat.

DAS MENTALE MODELL

Agenten erweitern Human Risk: ein Mitarbeiter, den niemand eingearbeitet hat, aber mit echten Zugangsdaten.

DAS VERHALTENSPROBLEM

Legitime KI und gekaperte Zugangsdaten erzeugen dasselbe Signal: Die Absicht steht nicht im Protokoll.

DIE NEUE EINSCHRÄNKUNG

Schaden summiert sich in API-Geschwindigkeit: der Wandel von einer alarmorientierten zu einer sanierungsorientierten Reaktion.

Agenten sind keine eigenständige Technologiekategorie neben Human Risk. Sie sind eine Erweiterung davon.

KI-Agenten werden von Menschen erschaffen. Sie handeln im Auftrag von Menschen. Sie tragen menschliche Zugangsdaten. Sie agieren in denselben Kommunikationskanälen wie E-Mail, Slack und Teams, in denen sich Human Risk bereits zeigt. Und ihre Fehlermodi wie Halluzination, Scope Creep und gegnerische Manipulation sind strukturell parallel zu den Fehlermodi, für die Human Risk Management ursprünglich entwickelt wurde.

Agenten sind Mitarbeitende, die Sie nie geschult haben.

Das nützlichste mentale Modell für einen KI-Agenten ist ein neuer Mitarbeiter, der nie eingearbeitet wurde. Er hat Zugriff, weil ihm jemand diesen Zugriff gegeben hat. Er wurde nicht darin geschult, was er mit diesem Zugriff tun darf. Er wird Fehler machen, genau wie Menschen Fehler machen, nur skalieren diese Fehler in der Geschwindigkeit der zugrunde liegenden Rechenleistung.

Diese Parallele ist nicht metaphorisch. Sie ist architektonisch. Die Kontrollen, die menschliches Verhalten steuern, wie Zugriffsrechte, Aktivitätsüberwachung, Verhaltensbaselines, Richtliniendurchsetzung und Anomalieerkennung, lassen sich direkt auf die Steuerung von Agentenverhalten übertragen. Unternehmen, die bereits eine ausgereifte Human-Risk-Praxis aufgebaut haben, sind der agentischen Bereitschaft näher, als ihnen vielleicht bewusst ist. Die Erweiterung ist kein Neuaufbau. Sie ist der Ausbau eines bestehenden Fundaments.

DAS MENTALE MODELL

Ein KI-Agent ist wie ein neuer Mitarbeiter, der nie eingearbeitet wurde: Er erhält Zugriff, ohne jemals geschult worden zu sein, wie er ihn nutzen soll, und kann jeden Fehler in Maschinengeschwindigkeit skalieren. Jede für den Menschen entwickelte Kontrolle lässt sich auf die Steuerung des Agenten übertragen.

Verhalten ist das Signal. Geschwindigkeit ist die neue Einschränkung.

Das Problem der Verhaltensmehrdeutigkeit

Eine der größten operativen Herausforderungen in der agentischen KI-Sicherheit ist die Unmöglichkeit, allein anhand des beobachtbaren Verhaltens zwischen legitimer KI-Aktivität und böswilliger Aktivität zu unterscheiden.

Eine Führungskraft, die einen KI-Agenten einsetzt, um in ihrem Namen Hunderte Slack-Nachrichten pro Tag zu versenden, ist aus Sicht der Verhaltensüberwachung nicht von einem Angreifer zu unterscheiden, der die Zugangsdaten dieser Führungskraft kompromittiert hat und sie für massenhaften Identitätsmissbrauch nutzt. Beide erzeugen dasselbe Signal: ein ungewöhnliches Nachrichtenvolumen von einer bekannten Identität. Das Einzige, was sie unterscheidet, ist die Absicht, und die Absicht ist im Ereignisprotokoll nicht sichtbar.

Die Lösung liegt im Verhaltenskontext: Baselines, die erfassen, wie „normal“ für einen bestimmten Nutzer aussieht, einschließlich seiner bekannten KI-Nutzungsmuster, sodass Abweichungen von dieser Baseline aussagekräftige Signale statt Rauschen erzeugen. Das ist keine neue Anforderung an Funktionalität. Es ist eine Erweiterung der Verhaltensintelligenz, die Human-Risk-Management-Plattformen bereits heute liefern.

Sanierung in Maschinengeschwindigkeit

Wenn ein KI-Agent entgleist, sei es durch Prompt Injection, Scope Creep, kompromittierte Zugangsdaten oder Fehlkonfiguration, summiert sich der Schaden nicht in menschlicher Geschwindigkeit. Er summiert sich in der Geschwindigkeit von API-Aufrufen.

Security-Operations, die nach dem Modell „erst Alarm, dann Untersuchung“ aufgebaut sind, sind auf diese Realität nicht kalibriert. Das neu entstehende Modell lautet „erst Sanierung, dann Alarm“: automatisierte Kontrollen, die eine Aktion stoppen, bevor eine menschliche Prüfung überhaupt möglich ist, wobei der Alarm als Nachweis dient, was abgefangen wurde, statt als Aufforderung zum menschlichen Eingreifen.

Das ist ein bedeutender architektonischer Wandel mit Auswirkungen auf die Plattformauswahl, das Workflow-Design und die SIEM-Integration.

DIE BEWERTUNGSFRAGE

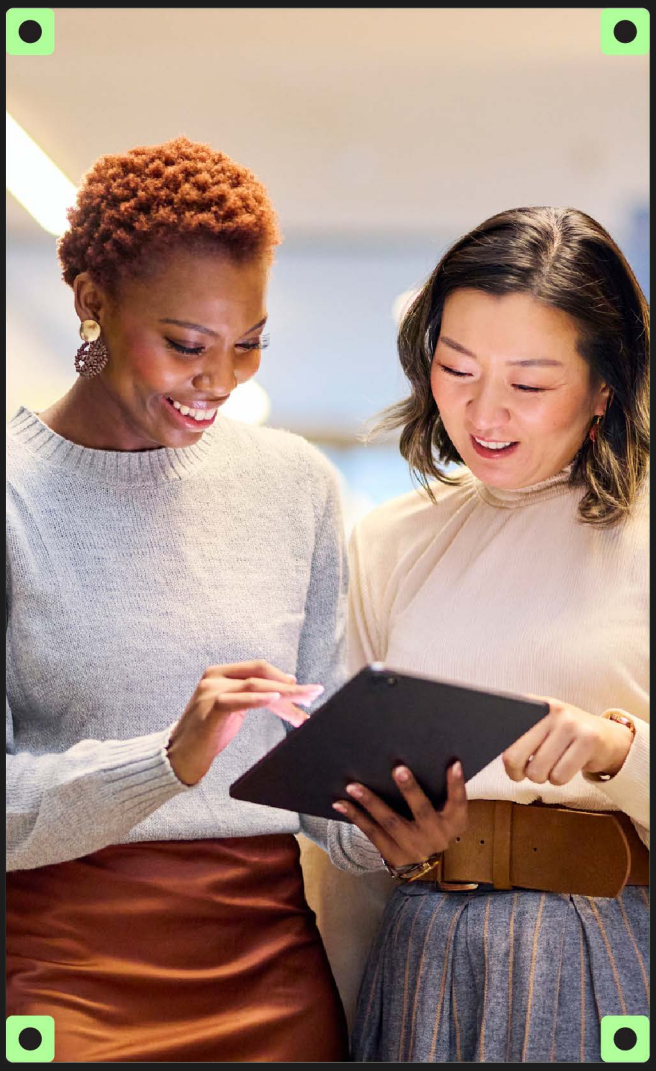
Sicherheitsverantwortliche, die 2026 Investitionen in KI-Sicherheit bewerten, sollten Anbieterfähigkeiten an einem konkreten Kriterium messen: Kann dieses Tool in der Geschwindigkeit der Bedrohung agieren?

DER FAHRPLAN

05.

Von Sichtbarkeit zu agentischer Bereitschaft.

Die meisten Unternehmen starten nicht bei null. Sie erweitern die E-Mail-Sicherheit, DLP und Identity Governance, die sie bereits betreiben. Dieser Abschnitt zeigt die vier Stufen der KI-Sicherheitsreife und das Gespräch mit dem Board, das ein Unternehmen dabei voranbringt.



Risikoinformierte Geschwindigkeit statt Risikoeliminierung: schnell bei KI vorangehen, aber den Moment kennen, in dem etwas schiefläuft.

DAS REIFEGRADMODELL

Vier Stufen, von Sichtbarkeit über Kontrolle und Governance bis zur agentischen Bereitschaft.

WIE SIE SICH EINSCHÄTZEN

Eine einzelne Diagnosefrage, die ein Unternehmen auf jeder Stufe verortet.

DAS GESPRÄCH MIT DEM BOARD

Drei Dinge, über die berichtet werden sollte, und die Darstellung, die die nächste Stufe finanziert.

Vier Stufen. Die meisten Unternehmen befinden sich zwischen Stufe eins und zwei.

Die meisten Unternehmen starten nicht bei null. Sie verfügen bereits über E-Mail-Sicherheit, Endpoint-Schutz, DLP und Identity Governance. Die Frage ist nicht, was von Grund auf neu aufgebaut werden muss. Sie lautet, wie bestehende Fähigkeiten erweitert, kalibriert und verknüpft werden können, um die in diesem Whitepaper beschriebene KI-spezifische Exposition zu adressieren.

STUFE	KERNFÄHIGKEITEN	SELBSTEINSCHÄTZUNG
01 / SICHTBARKEIT	**Den KI-Fußabdruck sehen.** Shadow-AI-Inventar · DLP-Überwachung der KI-Tool-Nutzung · Basis-Risiko-Scoring für Nutzer.	Können Sie benennen, welche KI-Tools heute aktiv sind? Sehen Sie, welche Daten durch sie fließen?
02 / KONTROLLE	**In Echtzeit durchsetzen.** Echtzeit-Blockierung von KI-Datenflüssen · Verhaltensbaselines je Nutzer · adaptive Richtliniendurchsetzung.	Können Sie Richtlinien in Echtzeit durchsetzen? Gelten Ihre Kontrollen sowohl für Agenten als auch für Menschen?
03 / GOVERNANCE	**Rechtssicher von Grund auf.** Forensischer Nachweis von KI-Interaktionen · LLM-Prompt-Protokollierung für eDiscovery · KI-Nutzungsrichtlinie mit messbarem Bekanntheitsgrad · rechtssichere Löschung von Daten nach Ablauf ihrer rechtlichen oder operativen Lebensdauer.	Können Sie KI-Prompt-Protokolle im Rahmen eines Legal Hold bereitstellen? Kennen mehr als 50 % Ihrer Mitarbeitenden Ihre KI-Richtlinie? Können Sie die aktive Löschung nicht mehr benötigter Daten nachweisen?
04 / AGENTISCHE BEREITSCHAFT	**Gebaut für Maschinengeschwindigkeit.** Katalogisierung von Agentenidentitäten · Erkennung von Verhaltensanomalien in Maschinengeschwindigkeit · sanierungsorientierte Architektur.	Haben Sie Ihre Human-Risk-Architektur auf KI-Agenten ausgeweitet? Können Ihre Tools in KI-Geschwindigkeit agieren?

2026 werden sich die meisten Unternehmen zwischen Stufe 1 und Stufe 2 wiederfinden. Kurzfristige Priorität ist es, die Lücke auf Stufe 1 zu schließen, also echte Sichtbarkeit in die Nutzung von KI-Tools und Datenflüsse herzustellen, bevor in fortgeschrittenere Kontrollebenen investiert wird. Governance und agentische Bereitschaft, die auf einem Sichtbarkeitsdefizit aufgebaut werden, erzeugen Richtlinien ohne Durchsetzung, das häufigste und teuerste Muster in der Unternehmens-KI-Sicherheit von heute.

Risikoinformierte Geschwindigkeit statt Risikoeliminierung.

Gespräche über KI-Sicherheit auf Board-Ebene haben 2026 einen bestimmten Charakter. Board-Mitglieder treiben gleichzeitig eine aggressive KI-Einführung für mehr Produktivität voran und verlangen von CISOs Rechenschaft über die dadurch entstehenden Risiken. Die Spannung ist real, und die Fragen werden nicht verschwinden.

Das wirksamste Board-Reporting zur KI-Sicherheit deckt drei Dinge ab: welche genehmigten KI-Tools im Einsatz sind und wie sie gesteuert werden, welche nicht genehmigte KI-Aktivität erkannt wurde und welche Kontrollen greifen, sowie den aktuellen Stand des Unternehmens bei den sechs Prinzipien dieses Frameworks.

Die richtige Darstellung für das Board ist nicht Risikoeliminierung, das ist weder erreichbar noch das richtige Ziel. Es ist risikoinformierte Geschwindigkeit: die Fähigkeit, bei der KI-Einführung schnell voranzugehen und dabei die Sichtbarkeit, Kontrollen und Governance-Architektur zu erhalten, um zu erkennen, wenn etwas schiefläuft, und in angemessener Geschwindigkeit zu reagieren.

GENEHMIGTE KI

Was im Einsatz ist, auf welche Daten jedes Tool zugreifen kann, wie es gesteuert wird und was sich seit dem letzten Bericht verändert hat.

NICHT GENEHMIGTE AKTIVITÄT

Was erkannt wurde, welche Kontrollen greifen und wo die verbleibende Exposition liegt.

STAND BEI DEN PRINZIPIEN

Ehrlicher aktueller Stand zu jedem der sechs Prinzipien, mit dem jeweils nächsten Schritt benannt.

DIE DARSTELLUNG

Risikoinformierte Geschwindigkeit ist das Ziel, nicht Risikoeliminierung. Die CISOs, die sich in diesem Umfeld bewähren, erfassen drei Dinge zugleich: Die menschliche Ebene bleibt die wichtigste Variable, Agenten sind eine Erweiterung davon, und Geschwindigkeit ist inzwischen eine Einschränkung, der Erkennung allein nicht gerecht werden kann.

FAZIT

06.

Das Zeitfenster ist jetzt.

Die Bedingungen, die 2026 zu einem entscheidenden Jahr für Investitionen in KI-Sicherheit machen, sind struktureller, nicht zyklischer Natur. Die Bedrohung ist real und nimmt zu. Die Bereitschaftslücke ist messbar und dokumentiert.

Sicherheitsarchitekturen, die nicht in vergleichbarer Geschwindigkeit reagieren können, sind strukturell im Nachteil, unabhängig davon, wie ausgereift ihre Erkennungsfähigkeiten sind.

Unternehmen, die KI-Sicherheit als eigenständige Initiative behandeln, als neue Budgetposition, neuen Anbieter oder neues Team, werden fragmentierte Abwehrmaßnahmen aufbauen. Unternehmen, die sie als natürliche Erweiterung ihrer bestehenden Human-Risk-Arbeit betrachten, werden etwas Kohärenteres, Verteidigungsfähigeres und Anpassungsfähigeres für noch unbenannte Bedrohungen aufbauen.

Die Entscheidungen, die Unternehmen in den nächsten zwölf bis achtzehn Monaten zu ihrer KI-Sicherheitsarchitektur treffen, werden sich deutlich schwerer rückgängig machen lassen, als sie beim ersten Mal richtig zu treffen.

Drei Dinge gelten gleichzeitig, und die CISOs, die sich in diesem Umfeld gut zurechtfinden, verstehen alle drei zugleich. Die menschliche Ebene bleibt die wichtigste Variable in der Unternehmenssicherheit, und KI hat sie komplexer gemacht, nicht einfacher. KI-Agenten sind keine eigenständige Sicherheitsdomäne. Sie sind eine Erweiterung von Human Risk, und die für Human Risk entwickelten Frameworks sind die richtige Grundlage, um Agentenrisiko zu adressieren. Und Geschwindigkeit ist eine neue Einschränkung. Die Bedrohung bewegt sich in Maschinengeschwindigkeit.

Sicherheitsverantwortliche, die ihre KI-Sicherheitsstrategie festlegen, bevor ein Vorfall das Gespräch erzwingt, werden in jeder Hinsicht besser aufgestellt sein: gegenüber ihrem Board, ihren Anbietern, ihren Aufsichtsbehörden und ihrem Unternehmen. Dieses Zeitfenster ist jetzt offen. Es wird nicht unbegrenzt offen bleiben.

Woher die Zahlen stammen.

Alle in diesem Whitepaper zitierten Datenpunkte mit Originalquellen.

STATISTIK	WERT	QUELLE
Unternehmen, die KI-gestützte Angriffe innerhalb von 12 Monaten erwarten	69 %	Mimecast State of Human Risk 2026
Unternehmen mit einer konkreten KI-Sicherheitsstrategie	40 %	Mimecast State of Human Risk 2026
Unternehmen mit nicht genehmigten KI-Tools	98 %	Mimecast Research
Mitarbeitende, die private GenAI-Konten bei der Arbeit nutzen	47 %	Netskope Cloud and Threat Report 2026
Mitarbeitende, die die KI-Nutzungsrichtlinie ihres Unternehmens kennen	18,5 %	Mimecast State of Human Risk 2026
KI-Interaktionen mit sensiblen Daten	39,7 %	Mimecast Research
KI-Prompts, die private Informationen preisgeben	4 %	Mimecast Research
An KI-Tools hochgeladene Dateien mit vertraulichem Inhalt	20 %	Mimecast Research
Tage bis zur Erkennung eines Shadow-AI-Vorfalls (Durchschnitt)	247	IBM 2025 Cost of a Data Breach Report
Mehrkosten eines Shadow-AI-Vorfalls gegenüber Standardvorfällen	670.000 US-Dollar	IBM 2025 Cost of a Data Breach Report
Ransomware-Vorfälle mit Beteiligung von KMU	88 %	Verizon DBIR 2025
Typische Kostenspanne für KMU-Sicherheitsvorfälle	120.000 bis 1,2 Mio. US-Dollar	Verizon DBIR 2025
Anteil von Phishing an allen Cyberangriffen	77 %	Mimecast GTI Report 2025
Anstieg KI-gestützter Betrugsfälle 2025	1.210 %	Pindrop Security AI Fraud Trends 2026
BEC-Angriffe, die heute KI-Deepfakes nutzen	40 %	Mimecast State of Human Risk 2026
Unternehmen mit einer KI-Nutzungsrichtlinie	44 %	Palo Alto Networks
Unternehmensanwendungen mit eingebetteten KI-Agenten bis 2026	40 % (Prognose)	Gartner

ANHANG

Quellen, Marken und Schlusswort.

QUELLEN

1. Mimecast 2025 Global Threat Intelligence Report.
2. Verizon 2025 Data Breach Investigations Report (DBIR).
3. NIST Center for AI Standards and Innovation (CAISI), Request for Information dated January 8, 2026.
4. Varonis 2025 State of Data Security Report.
5. Netskope Cloud and Threat Report 2026.
6. Mimecast State of Human Risk 2026 Report.
7. Harmonic Security 2025 AI Data Exposure Report.
8. IBM 2025 Cost of a Data Breach Report.
9. Palo Alto Networks, “6 Cybersecurity Predictions for the AI Economy in 2026,” Harvard Business Review, December 2025.

MARKEN

WEITERFÜHRENDE INHALTE

Mimecast-Research-Publikationen, einschließlich des zugrunde liegenden State of Human Risk 2026 Reports und der Global Threat Intelligence, sind verfügbar unter mimecast.com/resources.

WORK PROTECTED

Mimecast Sicherheit für Menschen, Daten & KI, damit Arbeit geschützt bleibt.

mimecast

Über diese Publikation.

Dieses Whitepaper dient ausschließlich Informationszwecken und stellt keine Rechts-, Regulierungs-, Finanz- oder sonstige Fachberatung dar. Mimecast hat angemessene Anstrengungen unternommen, um die Richtigkeit der in diesem Dokument enthaltenen Informationen sicherzustellen, gibt jedoch keine ausdrückliche oder stillschweigende Zusicherung oder Gewährleistung hinsichtlich ihrer Vollständigkeit oder Eignung für einen bestimmten Zweck.

Dieses Dokument spiegelt den Informationsstand zum Zeitpunkt der Veröffentlichung wider und kann ohne vorherige Ankündigung geändert werden. Leserinnen und Lesern wird empfohlen, unabhängigen fachlichen Rat einzuholen, bevor sie auf Grundlage der Inhalte dieses Whitepapers handeln.

Die Vervielfältigung oder Weiterverbreitung dieses Dokuments, ganz oder in wesentlichen Teilen, ohne vorherige schriftliche Zustimmung von Mimecast ist untersagt.

KONTAKT

Für Research-Anfragen, Briefings oder um eine individuelle Bewertung anhand des Reifegradmodells in diesem Whitepaper anzufordern: mimecast.com/company/contact/

WEITERFÜHRENDE INHALTE

Mimecast-Research-Publikationen, einschließlich des zugrunde liegenden State of Human Risk 2026 Reports und der Global Threat Intelligence, sind verfügbar unter mimecast.com/resources.

mimecast