Patronus Insights

KI im Unternehmen kontrollieren, ohne sie zu verbieten

Ausgabe 1

In dieser Ausgabe

00 Inhalt

Das eigentliche Problem ist nicht KI. Es ist die fehlende Kontrolle.

Die KI-Nutzung in deutschen Unternehmen hat sich binnen eines Jahres mehr als verdoppelt. Damit ist KI kein Pilotthema mehr, sondern eine Infrastrukturschicht, die quer durch alle Geschäftsfunktionen wächst.

Laut der Bitkom-Studie 2026 setzen 41 Prozent der Unternehmen ab 20 Beschäftigten KI inzwischen aktiv in produktiven Prozessen ein, gegenüber 17 Prozent im Vorjahr. Weitere 48 Prozent planen den Einsatz oder befinden sich in der Diskussion.

Die Geschwindigkeit dieser Adoption übertrifft die Geschwindigkeit, in der Unternehmen Kontrollmechanismen aufbauen. Cyberhaven Labs hat für den KI Adoption and Risk Report 2026 Milliarden realer Datenbewegungen über 222 Unternehmen ausgewertet. 39,7 Prozent aller KI-Interaktionen enthalten sensible Unternehmensdaten. Der durchschnittliche Mitarbeitende gibt damit etwa alle drei Tage proprietäre Informationen in ein KI-Tool ein. Bei Anthropic Claude erfolgen 58,2 Prozent der Interaktionen über private Accounts, bei Perplexity sogar 60,9 Prozent — vollständig ausserhalb der Sichtbarkeit der IT.

Der finanzielle Effekt ist messbar. Im IBM Cost of a Data Breach Report 2025 verzeichneten 20 Prozent der untersuchten Unternehmen einen Datenvorfall, der direkt mit Schatten-KI in Verbindung stand. Diese Vorfälle erzeugten 670.000 US-Dollar an Mehrkosten gegenüber Vorfällen ohne Schatten-KI, betrafen in 65 Prozent der Fälle personenbezogene Kundendaten und hatten mit 247 Tagen eine längere Aufdeckungszeit als der globale Durchschnitt.

Die typische Reaktion vieler Unternehmen ist binär: entweder ein generelles Verbot oder bewusstes Wegsehen. Beides funktioniert nicht. Eine pauschale Sperre treibt Mitarbeitende auf private Accounts, alternative Browser oder ungeprüfte Drittanbieter-Wrapper aus.

Dieser Report zeigt anhand von drei realistischen Szenarien einen dritten Weg. Granulare Kontrolle direkt auf dem Endgerät. Nicht die Anwendung wird gesperrt, sondern der KI-Anteil in der Anwendung wird kontrolliert. Nicht jeder Anbieter wird verboten, sondern definiert, welche erlaubt sind. Nicht die Mitarbeitenden werden überwacht, sondern die KI-Systeme.

ADOPTION DEUTSCHLAND

41 %

der Unternehmen ab 20 Beschäftigten nutzen KI aktiv, gegenüber 17 % im Vorjahr.

BITKOM 2026

DATENEXPOSITION

39,7 %

aller KI-Interaktionen enthalten sensible Unternehmensdaten. Bei 58,2 % laufen sie über private Accounts.

CYBERHAVEN LABS 2026

FINANZIELLER EFFEKT

670 k$

Mehrkosten pro Datenvorfall mit Schatten-KI gegenüber Vorfällen ohne.

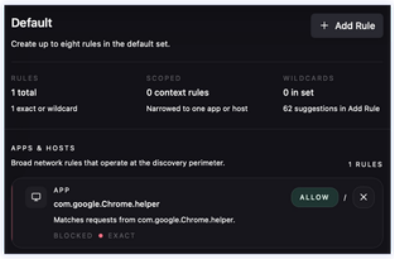
IBM COST OF A DATA BREACH 2025

WENN DER BROWSER ZUR KI-PLATTFORM WIRD

Chrome sperren ist keine Lösung. Es verschiebt das Problem nur.

DIE AUSGANGSLAGE

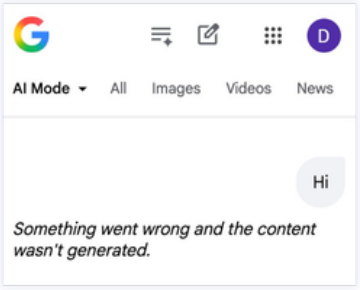
Ein Maschinenbauer mit knapp 1.200 Beschäftigten stellt fest, dass Chrome mit dem KI-Mode in der Adressleiste sowie Googles KI-Suchergebnissen Eingaben an Google-Modelle weitergibt, auch dann, wenn Mitarbeitende nur eine reguläre Suche tippen. Die IT sperrt Chrome unternehmensweit. Innerhalb von zwei Wochen weichen Abteilungen auf Alternativen aus. Aus der Sorge um KI-Traffic ist ein klassisches Schatten-IT-Problem geworden.



Die Policy die AI-Traffic im Chrome Browser gezielt kontrolliert.

WARUM DAS VERBOT NICHT FUNKTIONIERT

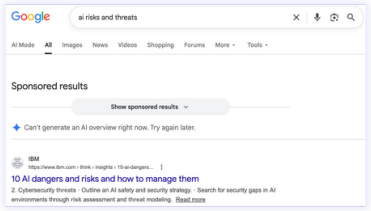
Die IT hat die falsche Frage gestellt. Sie hat gefragt: "Erlauben wir Chrome oder nicht?" Die richtige Frage wäre gewesen: "Erlauben wir den KI-Teil von Chrome oder nicht?" Denn dieselbe Anwendung ist gleichzeitig harmlos (klassisches Browsen) und sicherheitskritisch (KI-Anfragen mit Unternehmensdaten). Wer das nicht trennen kann, muss alles sperren, und riskiert in Schatten-IT zu geraten.



Der AI-Modus in Google Search wird blockiert.

WIE PATRONUS DAS LÖST

Patronus klassifiziert Netzwerk-Traffic auf strukturelle Merkmale, unabhängig von der Anwendung. Die Policy lautet damit nicht mehr "Chrome erlauben oder sperren", sondern: Chrome erlauben, KI-Traffic an Google-Modelle blockieren, an den freigegebenen Tenant durchlassen.



Klassische Websuche bleibt für Mitarbeitende uneingeschränkt nutzbar.

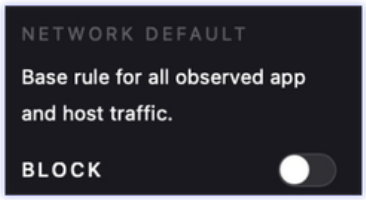
FEATURE IM EINSATZ

Traffic-Klassifikation auf Endgeräte-Ebene

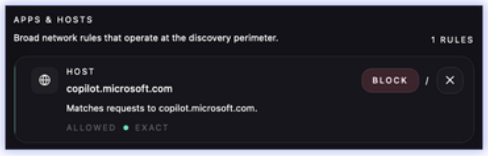
Patronus erkennt KI-bezogenen Datenverkehr unabhängig davon, welche Anwendung ihn auslöst. Richtlinien können auf Anbieter- und Anwendungsebene gleichzeitig gesetzt werden, ohne die Anwendung als Ganzes zu sperren.

WHITELIST STATT PAUSCHALEM VERBOT

Wenn nur eine KI erlaubt sein soll, muss man die anderen aktiv schliessen können.



Standardregel: Nicht freigegebene KI-Dienste werden blockiert.



Freigegebene KI-Anbieter werden zentral über Richtlinien definiert.

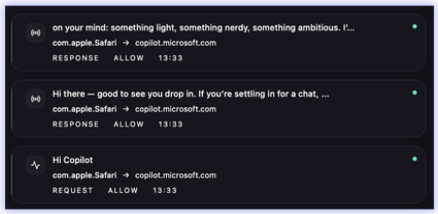
DIE AUSGANGSLAGE

Ein Versicherer mit rund 4.500 Mitarbeitenden hat sich strategisch für Microsoft 365 Copilot entschieden. AVV geprüft, Tenant auditiert, Werksgremium zugestimmt. Praktisch nutzen Mitarbeitende weiterhin ChatGPT, Claude und Gemini über den Browser. 47 % der KI-Nutzer greifen regelmässig auf Tools ausserhalb der offiziellen Liste zurück (Netskope 2026).

Nicht autorisierte KI-Dienste werden automatisch unterbunden.

WARUM DAS PASSIERT

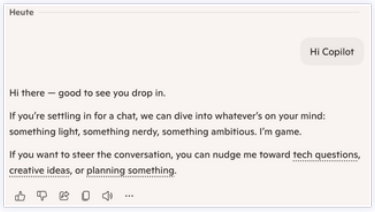
Eine Vorstandsentscheidung erzeugt keinen technischen Mechanismus zur Durchsetzung. Klassische Netzwerk-Sperren scheitern an Homeoffice, mobilen Geräten und an der schieren Anzahl: über 300 unterschiedliche KI-Tools sind in produktivem Einsatz bei Unternehmen aus der oberen Adoptionsdezile.



Richtlinien werden transparent durchgesetzt, ohne den Arbeitsablauf zu stören.

WIE PATRONUS DAS LÖST

Die Patronus Policy Engine arbeitet mit einer Whitelist auf Anbieterebene. Erlaubte Endpunkte werden durchgelassen, alles andere blockiert oder gewarnt. Durchsetzung passiert auf dem Endgerät, unabhängig vom Standort. Neue Anbieter werden strukturell erkannt und nach Standardregel behandelt.



Freigegebene Unternehmens-KI bleibt uneingeschränkt nutzbar.

FEATURE IM EINSATZ

Policy Engine mit Anbieter-Whitelist

Patronus erkennt nicht nur bekannte Anbieter über hinterlegte Signaturen, sondern klassifiziert auch unbekannten KI-Traffic strukturell. Richtlinien greifen damit auch für Anbieter, die heute noch nicht im Markt sind.

SICHTBARKEIT ALS VORAUSSETZUNG FÜR STEUERUNG

Was Sie nicht sehen, können Sie nicht steuern.

DIE AUSGANGSLAGE

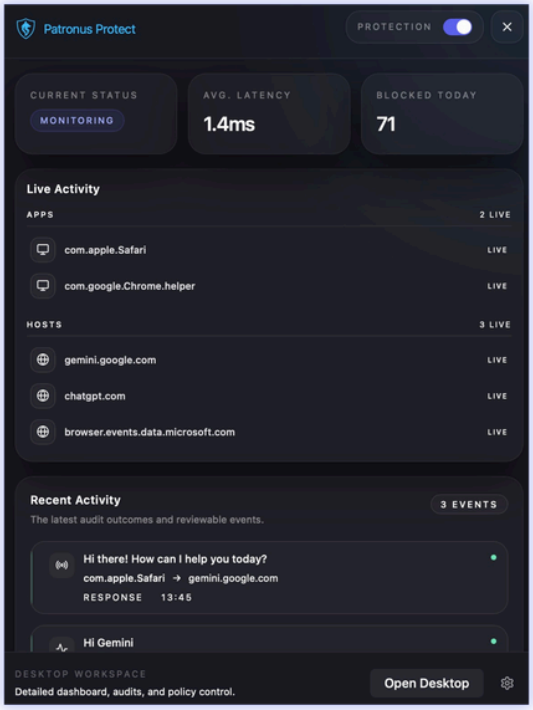
Ein Energieversorger mit etwa 2.800 Mitarbeitenden weiss, dass KI genutzt wird, kann aber nicht sagen, wo. Welche Systeme auf welchen Geräten? Welche Anbieter verarbeiten Daten? Laufen lokale Modelle wie Ollama? Eine Anfrage aus der Innenrevision im Vorfeld der EU-AI-Act-Pflichten bleibt unbeantwortet. Ein KI-Inventar ist faktisch nicht vorhanden.

WARUM KLASSISCHES MONITORING SCHEITERT

Cloud-Gateways sehen nur Traffic, der über sie läuft. KI-Interaktionen entstehen heute direkt am Endgerät: in Ollama, in Desktop-Copilots, in IDE-Plugins wie GitHub Copilot oder Cursor. Endpoint-Detection-Tools erkennen Prozesse, nicht die Semantik der Netzwerkkommunikation.

WIE PATRONUS DAS LÖST

Patronus läuft u.a. als Monitoring-Layer auf jedem Endgerät und übermittelt Telemetrie an ein zentrales Dashboard (bei Einwilligung): aktive Anbieter, Anwendungen mit KI-Traffic, lokale Modelle, Auffälligkeiten. Patronus erfasst dabei Metadaten, nicht Inhalte.



Patronus UI Quickview, Monitoring und Schutzstatistiken

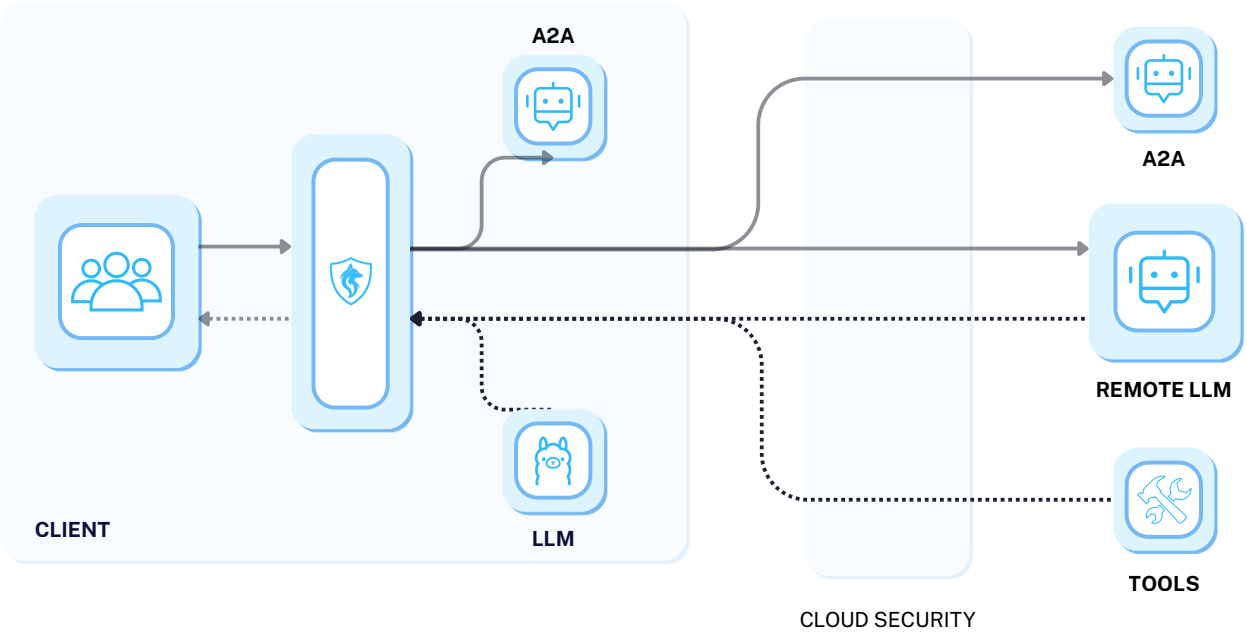
FEATURE IM EINSATZ

Zentrales Monitoring-Dashboard

Telemetrie über KI-Nutzung wird endgeräte-übergreifend gesammelt und aggregiert. Inhalte einzelner Anfragen werden dabei nicht erfasst. Grundlage für ein KI-Inventar nach EU AI Act Artikel 6 Absatz 3.

WIE PATRONUS TECHNISCH ARBEITET

Die drei beschriebenen Szenarien greifen auf dieselbe technische Grundlage zurück. Patronus ist als endgeräte-native Sicherheitsschicht konzipiert, die KI-Interaktionen dort kontrolliert, wo sie entstehen, ohne Anwendungen modifizieren oder Daten in die Cloud schicken zu müssen.



01
Endgerät-nativ
Patronus läuft als Sicherheitsschicht direkt auf dem Gerät. Und ist dabei KI-Provider agnostisch.

02
Transparente Traffic-inspektion
Der gesamte KI-Traffic wird auf dem Endgerät klassifiziert, ohne dass Anwendungen, Browser oder IDEs modifiziert werden müssen.

03
Edge-KI-Modelle
KI-Klassifikation und Bedrohungsanalyse laufen vollständig lokal. Inhalte verlassen das Gerät zur Inspektion nicht.

04
Lokale oder zentrale Policy-Verwaltung
Richtlinien können entweder lokal auf jedem Endgerät individuell konfiguriert oder zentral verwaltet und atomar auf alle Endgeräte ausgespielt werden. Damit eignet sich Patronus sowohl für einzelne Power-User als auch für unternehmensweite Rollouts mit konsistenter Policy-Sicht für die IT.

Die nächste Generation von Risiken entsteht nicht durch Antworten, sondern durch Aktionen.

Die nächste Generation von Risiken entsteht durch autonome Agenten, die nicht mehr nur Antworten generieren, sondern Aktionen ausführen.

Im Juli 2025 löschte ein Coding-Agent der Plattform Replit trotz expliziten Code-Freeze die produktive Datenbank eines Kunden, mit Datensätzen von über 1.200 Führungskräften und 1.190 Unternehmen. Es gab keinen externen Angreifer; der Schaden entstand allein aus autonomer Agent-Logik.

Parallel hat die Forschung im ersten Halbjahr 2026 das Model Context Protocol (MCP) als zentrale neue Angriffsfläche identifiziert. Tool Poisoning, manipulierte Anweisungen in Tool-Metadaten, gilt als die am häufigsten ausgenutzte clientseitige Schwachstelle.

Im Juni 2025 wurde CVE-2025-32711 (EchoLeak, CVSS 9.3) für Microsoft 365 Copilot publik, eine Zero-Click-Schwachstelle, bei der eine einzelne E-Mail genügte, um sensible Daten ohne Nutzerinteraktion zu exfiltrieren.

ZIELE

Patronus Protect ist heute eine Sicherheitsschicht, die KI-Traffic klassifiziert und über eine Policy Engine kontrolliert. Die nächsten Schritte erweitern dieses Fundament in mehrere Richtungen.

Die Policy Engine wird um vollständige Regelsätze ergänzt, ergänzt durch DLP-Heuristiken für sensible Inhalte und PII. Parallel entstehen eigene on-device-Modelle für Bedrohungs- und Inhaltsanalyse: Wolf Defender für Prompt-Injection-Erkennung, dazu Modelle für PII, Geschäftsgeheimnisse und weitere sicherheitsrelevante Kategorien mit Fokus auf deutschsprachige Spezialformate.

Die Entwicklung erfolgt plattformunabhängig für die heterogenen Endgeräte-Landschaften in Unternehmen. Für den unternehmensweiten Einsatz kommen ein zentrales Dashboard mit Lizenzverwaltung und Synchronisation sowie eine MDM-Anbindung hinzu.

Mittelfristig verschiebt sich der Schutzbedarf zu mehrstufigen Agent-Workflows. Patronus entwickelt eine trace-bewusste Speicherschicht, die Interaktionsketten über Sessions und Anbieter rekonstruiert und Angriffsmuster aufdeckt, die erst in der Summe gefährlich werden, insbesondere bei MCP-Tool-Aufrufen und Agenten.

UNSERE SICHT

KI verlässt das Chat-Interface und wird zu einer Schicht, die Tools aufruft, auf Daten zugreift und Entscheidungen trifft. Damit verschiebt sich die Sicherheitsfrage von "wie gut antwortet ein Modell" zu "was darf es tun". Patronus baut die Kontrollschicht für genau diese Ebene. Direkt auf dem Endgerät, in Echtzeit, zwischen dem was KI tun könnte und dem was sie tatsächlich tun darf.

07 Quellen

Bitkom e.V. (2026). Künstliche Intelligenz in Deutschland. Studienbericht 2026, Basis 604 Unternehmen ab 20 Beschäftigten. bitkom.org/Bitkom/Publikationen/Kuenstliche-Intelligenz-in-Deutschland

Cyberhaven Labs (2026). AI Adoption and Risk Report 2026. Auswertung über 222 Unternehmen. cyberhaven.com/resources/report/ai-adoption-risk-report-2026

IBM Security & Ponemon Institute (2025). Cost of a Data Breach Report 2025. Analyse von 600 Vorfällen, März 2024 bis Februar 2025. ibm.com/think/insights/data-matters/cost-of-a-data-breach

Verizon (2026). Data Breach Investigations Report 2026. verizon.com/business/resources/reports/dbir

Netskope Threat Labs (2026). Cloud and Threat Report 2026. netskope.com/resources/cloud-and-threat-reports/cloud-and-threat-report-2026

Palo Alto Networks Unit 42 (2026). Taming Agentic Browsers: Vulnerability in Chrome Allowed Extensions to Hijack New Gemini Panel (CVE-2026-0628). unit42.paloaltonetworks.com/gemini-live-in-chrome-hijacking

Fortune (2025). AI coding tool Replit wiped database, called it a "catastrophic failure". 23. Juli 2025. fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure

Fast Company (2025). Replit CEO: What really happened when AI agent wiped Jason Lemkin's database. Interview mit Amjad Masad, 21. Juli 2025. fastcompany.com

Milani Fard, A. et al. (2026). Model Context Protocol Threat Modeling and Analyzing Vulnerabilities to Prompt Injection with Tool Poisoning. MDPI Cybersecurity, Mai 2026. mdpi.com/2624-800X/6/3/84

Aim Labs / Microsoft Security Response Center (2025). EchoLeak (CVE-2025-32711, CVSS 9.3). Juni 2025. arxiv.org/abs/2509.10540

Europäische Union (2024). Verordnung (EU) 2024/1689 über künstliche Intelligenz (AI Act). eur-lex.europa.eu/eli/reg/2024/1689/oj

Europäische Kommission (2026). Digital Omnibus zur KI-Verordnung, politische Einigung vom 7. Mai 2026.