Patronus Radar

AI Security Briefing

Drei Monate AI Security im Überblick

Ausgabe 1

Über diesen Report

Patronus Radar erscheint monatlich und ordnet die wichtigsten Entwicklungen der AI Security für IT-Verantwortliche und Entscheider im DACH-Raum ein. Herausgeber ist Casdo Labs. Mit Patronus Protect entwickelt Casdo Labs eine On-Device-Sicherheitsschicht, die KI- und Agenten-Traffic direkt auf dem Endgerät erfasst, klassifiziert und vor der Ausführung kontrolliert. Über eigene Policies legen Unternehmen selbst fest, was erlaubt ist und was nicht, unabhängig vom KI-Anbieter.

Mehr unter patronus.studio.

Rückfragen, Hinweise und Themenvorschläge
für die nächste Ausgabe: team@patronus.studio

In dieser Ausgabe

00 Inhalt

Hinter den Vorfällen des Quartals stehen drei strukturelle Risiken.

Q2 2026 hat eine Vielzahl konkreter KI-Sicherheitsvorfälle gebracht, von Supply-Chain-Angriffen über Hugging Face bis zu Zero-Click-Vektoren in KI-Browsern. Hinter den Einzelfällen stehen drei Muster, die jedes Unternehmen mit KI-Einsatz in den nächsten Monaten adressieren sollte. Dieser Report ordnet sie ein.

01 Compromised KI Tooling

Kategorie: Supply Chain

Was ist das?

Angriffe nicht auf die KI selbst, sondern auf die Werkzeuge, mit denen Entwickler und Unternehmen KI in ihre Systeme integrieren. Beispiele sind Open-Source-Bibliotheken, Modell-Repositorys und Protokolle wie MCP, über die KI-Anwendungen mit anderen Tools sprechen.

Was Q2 gezeigt hat

In Q2 wurden architektonische Schwachstellen im Model Context Protocol publik, ein Supply-Chain-Vorfall in der Open-Source-Bibliothek LiteLLM zog tausende Unternehmen in Mitleidenschaft, und auf Hugging Face wurden fast 500 Modelle mit eingebetteter Malware identifiziert.
Wer KI-Tooling einsetzt, ohne dessen Lieferkette zu prüfen, akzeptiert ein blindes Vertrauensverhältnis bis tief in den Stack.

02 Agent Hijacking

Kategorie: Operational Compromise

Was ist das?

KI-Agenten sind Systeme, die nicht nur Antworten generieren, sondern Aktionen ausführen, etwa Dateien lesen, E-Mails verschicken, Code ausführen oder Tools aufrufen. Agent Hijacking bezeichnet Angriffe, bei denen ein Angreifer diese Aktionen über manipulierte Inhalte fernsteuert, ohne dass der eigentliche Nutzer etwas davon merkt.

Was Q2 gezeigt hat

KI-Agenten von Coding-Tools bis zu KI-Browsern führten Aktionen aus, die der Nutzer nicht autorisiert hatte. SymJack, TrustFall, CometJacking und der erste dokumentierte LLM-Agent-Intrusion-Fall von Sysdig zeigen, dass indirekte Prompt Injection nicht mehr theoretisch ist. Zero-Click-Vektoren reichen von Kalendereinladungen über Webseiten bis zu kompromittierten MCP-Servern.

03 Shadow KI als Compliance-Problem

Kategorie: Governance / DACH-spezifisch

Was ist das?

Shadow KI bezeichnet die Nutzung von KI-Tools im Unternehmen, ohne dass die IT-Abteilung davon weiß oder sie freigegeben hat. Typisches Beispiel: Mitarbeitende nutzen ChatGPT, Claude oder Copilot über private Accounts, weil die offiziell freigegebenen Tools langsamer oder umständlicher sind.

Was Q2 gezeigt hat

77 Prozent der KI-aktiven Unternehmen mit SAP-Stack nutzen non-SAP-Tools (DSAG 2026). 31 Prozent der Organisationen können nicht sagen, ob sie überhaupt einen agentischen Breach hatten (HiddenLayer 2026). Das sind keine Awareness-Lücken mehr, sondern strukturelle Sichtbarkeits-Probleme, die DORA, NIS2 und der EU AI Act in den nächsten Monaten verschärfen werden.

WENN DER HERSTELLER DAS RISIKO ABSCHIEBT

Anthropic nennt es erwartetes Verhalten. 200.000 MCP-Server nennen es Angriffsfläche.

DIE AUSGANGSLAGE

Am 15. April 2026 veröffentlichte OX Security eine Analyse, wonach ein Designfehler im Model Context Protocol Arbitrary Command Execution auf jedem System mit verwundbarer MCP-Implementierung ermöglicht. Betroffen sind die offiziellen SDKs in Python, TypeScript, Java und Rust. Anthropic stufte das Verhalten als erwartet ein und verzichtet auf eine Protokolländerung. The Register beziffert die Exposition auf rund 200.000 Server, OX selbst auf über 7.000 Server mit insgesamt mehr als 150 Millionen Downloads. In derselben Disclosure dokumentierte OX zehn weitere CVEs in MCP-STDIO-Konfigurationen, darunter ein Zero-Click-Vektor in Windsurf 1.9544.26, der über manipulierten HTML-Inhalt die unbemerkte Registrierung eines bösartigen MCP-Servers erlaubt.

EXPOSITION

200.000

MCP-Server mit verwundbarer SDK-Konfiguration laut The Register

THE REGISTER 2026

DOWNLOADS

150M+

Downloads betroffener SDKs in Python, TypeScript, Java, Rust

OX SECURITY 2026

WARUM DIE STANDARDMITIGATIONEN NICHT GREIFEN

Die typische Empfehlung lautet, nur vertrauenswürdige MCP-Server zu installieren. Das hilft nicht, wenn der Fehler in der Protokoll-Architektur selbst sitzt. Wer Claude Code, Cursor oder GitHub Copilot im Unternehmen einsetzt, hat MCP irgendwo im Stack, oft ohne dies aktiv konfiguriert zu haben. Klassische Domain-Sperren auf Netzwerk-Ebene greifen nicht, weil der Tool-Aufruf innerhalb des regulären Requests an den KI-Provider liegt. Cloud-Gateways sehen den Vektor nicht, weil er strukturell wie legitimer KI-Traffic aussieht.

DISCLOSURE

10 CVEs

in einer einzigen Veröffentlichung am 15. April 2026

OX SECURITY 2026

WAS DAS FÜR UNTERNEHMEN BEDEUTET

Solange Anthropic, Google und Microsoft ähnliche Vektoren als bekanntes Verhalten klassifizieren, verlagert sich die Sicherheits-Verantwortung zum Anwender. Wirksamer Schutz setzt auf Endgeräte-Ebene an, dort wo MCP-Aufrufe tatsächlich ausgeführt werden, und kontrolliert sie vor der Ausführung. Eine reine Provider-Whitelist reicht nicht, weil die Gefahr in der Anfrage selbst liegt, nicht im Ziel.

WENN DAS TOOL ZUM ANGREIFER WIRD

Ein Klick reicht. Manchmal nicht einmal das.

SymJack
Symlink-RCE
6 Coding Agents
06/2026

TrustFall
One-Click-RCE
Claude Code, Cursor, Gemini CLI, Copilot
06/2026

PerplexedComet
Zero-Click via Calendar Invite
03/2026

CometJacking
Browser-Hijack via Indirect Injection
06/2026

Sysdig
Erste LLM-Agent-Intrusion im Feld
05/2026

DIE AUSGANGSLAGE

Adversa AI meldete im Juni 2026 zwei Exploits, die einen großen Teil des KI-Coding-Stacks gleichzeitig betreffen. SymJack ist ein Symlink-Hijack-RCE gegen sechs KI-Coding-Agenten. TrustFall ist ein One-Click-RCE gegen Claude Code, Cursor, Gemini CLI und GitHub Copilot über einen regressierten Trust-Dialog. Parallel führte Microsoft Prompt Injections bis zu host-level RCE in Semantic Kernel zurück, und ein DEF-CON-Talk verkettete eine indirekte Injection zu einem persistenten Copilot-Backdoor.

WENN DER BROWSER ZUM AGENTEN WIRD

Zenity Labs dokumentierte mit PerplexedComet einen Zero-Click-Vektor gegen Perplexitys Comet-Browser. Eine Kalendereinladung mit eingebetteten Instruktionen genügt, damit der Agent das lokale Dateisystem liest und Inhalte exfiltriert. LayerX zeigte mit CometJacking, wie sich derselbe Browser über einen einzigen Klick gegen den Nutzer wenden lässt, inklusive Übernahme von Password-Managern. Perplexity benötigte bis zum finalen Patch 120 Tage.

DIE ERSTE LLM-AGENT-INTRUSION IM FELD

Am 10. Mai 2026 dokumentierte Sysdig eine Intrusion, bei der ein LLM-Agent eine Vier-Pivot-Kompromittierung bis zur Datenbank-Exfiltration in unter einer Stunde durchführte, ohne menschliche Steuerung. Einstieg war CVE-2026-39987, eine Pre-Auth-RCE im Python-Notebook Marimo. Es ist der erste dokumentierte Fall eines autonom agierenden Angreifer-Agenten.

WAS DAS FÜR UNTERNEHMEN BEDEUTET

So unterschiedlich die Vektoren sind, der gemeinsame Nenner ist derselbe: die schädliche Aktion wird auf dem Endgerät ausgeführt. Kontrolle am Netzwerk-Perimeter kommt zu spät, weil der Traffic dort wie legitime KI-Nutzung aussieht. Wirksam ist nur, was Agent-Aktionen am Punkt der Ausführung sichtbar macht und vor der Ausführung bewertet.

BEOBACHTUNG AM RAND

University of Toronto, 2. Juni 2026. Ein Team um Nicolas Papernot publizierte einen Proof-of-Concept für einen KI-getriebenen Wurm, der seine Angriffsstrategie adaptiv an das jeweilige Ziel anpasst. In Tests kompromittierte er 73,8 Prozent eines simulierten 33-Host-Netzwerks in sieben Tagen. Noch sind das Laborbedingungen, aber autonome KI-Angriffe sind kein theoretisches Risiko mehr.

SICHTBARKEIT ALS VORAUSSETZUNG FÜR COMPLIANCE

77 Prozent der SAP-Anwender nutzen non-SAP-KI. Das ist kein Mitarbeiterproblem.

DIE AUSGANGSLAGE

Die DSAG veröffentlichte im Frühjahr 2026 ihre jährliche Investment Survey. Nur 3 Prozent der SAP-Kunden nutzen SAP Business AI produktiv, aber 77 Prozent der KI-aktiven Unternehmen arbeiten mit Tools wie Copilot, ChatGPT und Claude, häufig über persönliche Accounts und ohne formale Freigabe. Das ist Shadow KI in einer Industrie, die durch DORA, NIS2 und den EU AI Act unter besonderer Dokumentationspflicht steht.

WARUM DAS KEINE AWARENESS-LÜCKE IST

Laut dem 2026 AI Threat Landscape Report von HiddenLayer geht eine von acht gemeldeten KI-Sicherheitsverletzungen auf agentische Systeme zurück. 31 Prozent der Organisationen können nicht sagen, ob sie überhaupt einen agentischen Breach hatten. Das ist kein Bewusstseinsproblem, sondern ein Architekturproblem: klassische DLP- und Netzwerk-Tools sehen diese Interaktionen nicht, weil sie über autorisierte Provider laufen oder lokal auf Endgeräten stattfinden.

DER VERCEL-VORFALL ALS LEHRSTÜCK

Im April 2026 wurde bekannt, dass ein Vercel-Mitarbeiter dem KI-Tool Context.ai per OAuth Vollzugriff auf seinen Google Workspace gegeben hatte. Context.ai war zwei Monate zuvor über Lumma-Stealer-Malware kompromittiert worden. Die exfiltrierten Tokens nutzten Angreifer für laterale Bewegung in Vercels Systeme. Eine einzige OAuth-Zustimmung reichte aus.

WAS UNTERNEHMEN JETZT BRAUCHEN

Was fehlt, ist ein Inventar der tatsächlich aktiven KI-Tools und ihrer Datenflüsse. Solange unklar ist, welche Systeme auf welchen Endgeräten laufen und welche Berechtigungen erteilt wurden, ist keine Compliance-Aussage gegenüber Aufsichtsbehörden belegbar. Erfassen lässt sich das nur auf dem Endgerät selbst. Dort laufen alle KI-Interaktionen zusammen, freigegeben oder nicht.

SHADOW KI

77 %

der KI-aktiven SAP-Kunden nutzen non-SAP-Tools wie Copilot, ChatGPT oder Claude

DSAG 2026

AGENTIC BREACHES

1/8

aller KI-Breaches geht auf agentische Systeme zurück · HIDDENLAYER 2026

HIDDENLAYER 2026

SICHTBARKEITS-LÜCKE

31 %

der Unternehmen wissen nicht, ob sie einen agentischen Breach hatten

HIDDENLAYER 2026

WAS SICH FÜR EUROPÄISCHE UNTERNEHMEN ÄNDERT

Digital Omnibus schafft Zeit, nicht Entlastung.

DER STATUS DES DIGITAL OMNIBUS

Am 7. Mai 2026 erzielten Europäisches Parlament und Rat eine politische Einigung zum sogenannten Digital Omnibus. Die Hochrisiko-Pflichten des AI Acts würden damit von August 2026 auf Dezember 2027 verschoben. Die formale Annahme steht zum Redaktionsschluss noch aus. Unstrittig ist: die GPAI-Pflichten gelten seit August 2025, die vollen Durchsetzungsbefugnisse des AI Office einschließlich Bußgelder aktivieren sich planmäßig im August 2026, und DSGVO und AI Act gelten parallel.

WAS UNTERNEHMEN JETZT TUN SOLLTEN

Die zusätzliche Zeit ist kein Grund zur Verzögerung. Die eigentliche Arbeit ist nicht das Ausfüllen von Templates, sondern der Aufbau eines vollständigen KI-Inventars und die Klassifizierung nach Anhang III. Ein solches Inventar entsteht nicht per Self-Reporting, sondern nur durch technische Erfassung auf Endgeräte-Ebene. Wer jetzt anfängt, hat 18 Monate für Verfeinerung. Wer Ende 2027 anfängt, hat Wochen.

DSGVO-BUSSGELDER ALS REALITÄTSCHECK

Zwischen Januar 2025 und Januar 2026 verhängten europäische Behörden DSGVO-Bußgelder von rund 1,2 Milliarden Euro. Ein erheblicher Teil entstand nicht durch den Vorfall selbst, sondern durch unvollständige Meldungen nach Artikel 33. Free Mobile in Frankreich zahlte allein 8 Millionen Euro für eine unvollständige Vorfallsdokumentation. Wer KI-Vorfälle nicht erkennt, kann sie nicht fristgerecht melden.

BSI UND NATIONALE AUFSICHT

Das BSI aktualisierte im Quartal seine Leitlinien zu Evasion-Angriffen auf LLMs und publizierte gemeinsam mit der französischen ANSSI Empfehlungen zum sicheren Einsatz von KI-Coding-Assistenten. Mit dem deutschen KI-Maßnahmen- und Innovationsgesetz erhält die BaFin ein breites Mandat zur Aufsicht über KI-Systeme im regulierten Finanzbereich, in Abstimmung mit BNetzA und BSI. Mit der NIS2-Umsetzung seit Dezember 2025 wächst die Schnittmenge zwischen Cybersecurity- und KI-Pflichten weiter.

- GPAI-Pflichten in Kraft — **02.08.2025**
- EU-Parlament verabschiedet Verhandlungsposition (569:45) — **26.03.2026**
- Vorläufige Trilog-Einigung — **07.05.2026**
- IMCO/LIBE bestätigen Einigung — **02.06.2026**
- Transparenzpflichten Art. 50 Durchsetzungs-befugnisse AI Office — **02.08.2026**
- Wasserzeichen-Pflicht — **02.12.2026**
- Hochrisiko-Pflichten Anhang III (sofern Omnibus verabschiedet) — **02.12.2027**

06 Quellen

Studien und Reports

HiddenLayer (2026). 2026 AI Threat Landscape Report. Befragung von 250 IT- und Security-Leadern. hiddenlayer.com

DSAG (2026). DSAG Investment Survey 2026. Deutschsprachige SAP-Anwendergruppe. dsag.de

OX Security (2026). The Mother of All AI Supply Chains: Critical Systemic Vulnerability at the Core of the MCP. 15. April 2026. ox.security

JFrog (2026). Software Supply Chain Security State of the Union Report 2026. jfrog.com
Vorfälle und Disclosures

OX Security (2026). MCP Supply Chain Advisory: RCE Vulnerabilities Across the AI Ecosystem. ox.security

The Register (2026). Anthropic MCP design flaw exposes 200,000 servers. 16. April 2026. theregister.com

Adversa AI (2026). Top Agentic AI Security Resources, June 2026. SymJack und TrustFall Disclosures. adversa.ai

Zenity Labs (2026). PerplexedComet: Zero-Click Browser Agent Compromise. März 2026.

LayerX (2026). CometJacking: How One Click Can Turn Perplexity's Comet AI Browser Against You. layerxsecurity.com

Sysdig (2026). First Documented LLM Agent Intrusion in the Wild. 10. Mai 2026.

BlueRadius Cyber (2026). AI Cybersecurity Incident Report 2026: Vercel and Context.ai.

University of Toronto / CleverHans Lab (2026). AI Agents Enable Adaptive Computer Worms. arXiv, 2. Juni 2026.

Regulierung

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 7. Mai 2026.

BSI. Evasion-Angriffe auf LLMs – Gegenmaßnahmen in der Praxis. bsi.bund.de

BSI / ANSSI. Empfehlungen zum sicheren Einsatz von KI-Coding-Assistenten.

Bundesregierung (2026). KI-Maßnahmen- und Innovationsgesetz (KI-MIG).