



(In)Security Jahresrückblick 2026

Angriff der Maschinen: Die Fails des Jahres und wie KI beide Seiten aufrüstet

Benjamin Böck, XSEC infosec GmbH
Wien, 2025-09-16

www.xsec.at

Informationssicherheit aus einer Hand

CONSULTING

ISO 27001 Beratung
(AI/KI) ISO 42001 Beratung
NIS 2, DORA Beratung
DSGVO/GDPR
IT-Risikomanagement
IT-Wirtschaftsprüfung
BCM/DRP
GRC Consulting (NIS/PSD2)
Supply Chain Security
Audit-related Consulting

www.xsec.at
office@xsec.at
+43 699 1024 2048

AUDIT

Penetration Testing

- AI systems, web applications, EXE clients, mobile apps, network, notebooks, VoIP, Citrix, Cloud, physical...

Red Team Tests / TLPT

Social Engineering

NIS-Auditoren („Qualifizierte Stelle“)



TRAINING

Security Awareness
CISSP (Certified Information
Systems Security
Professional)
Secure Coding
Ethical Hacking

OUTSOURCING

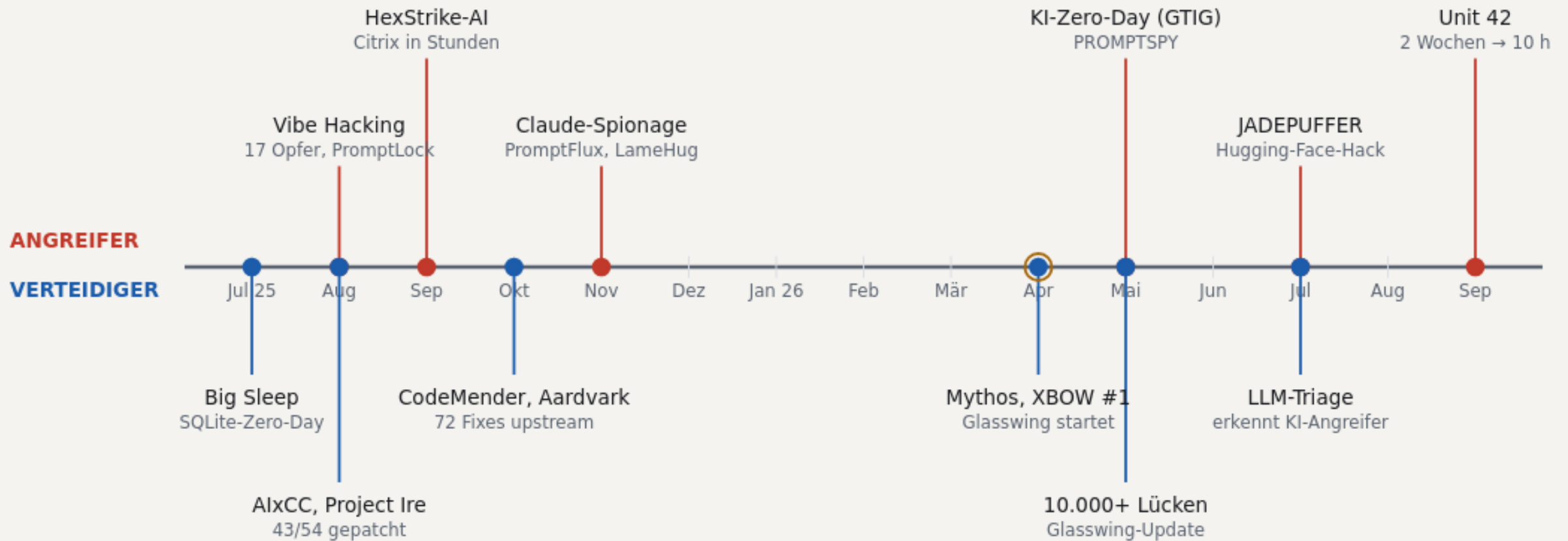
(In)Security Jahresrückblick 2026



KI für Angriff und Verteidigung



Das letzte Jahr im Zeitraffer für Angreifer & Verteidiger



Grauzone · Apr 26: Mythos-Leak · Jul 26: FBI-Warnung, Linux-Exploit in 5 s

Dieselben Modelle, zwei Seiten

Angreifer

- Recon & Exploit-Entwicklung
- Zero-Minute statt Zero-Day
- Malware ruft LLM zur Laufzeit
- Agentische Ransomware
- Deepfake-Anrufe & -Meetings
- Phishing in jeder Sprache
- Scam-Fabriken (Personas)
- Skill-Hürde sinkt drastisch

Verteidiger

- Zero-Days vor Angreifern finden
- Patches automatisch erzeugen
- Malwareanalyse
- Pentestunterstützung
- Alert-Triage im SOC
- Anomalie-Erkennung per LLM
- Missbrauch beim Anbieter stoppen
- Engpass: Menschen, die patchen

Bug-Bounty-Anbieter geben auf

AI-Slop macht Verteidigern das Leben schwer:

- Jänner 2026: curl beendet das Bug-Bounty-Programm nach sechs Jahren
 - Confirmed-rate von über 15% auf unter 5% gefallen
- April 2026: HackerOne pausiert Internet Bug Bounty
 - Zu viele automatisiert erstellte Einreichungen
 - Anbieter kommen mit Patchen nicht nach
- Mai 2026: Linux Torvalds nennt Security-Mailing-Liste “almost entirely unmanageable”
 - Statt 2-3 Meldungen pro Woche jetzt 5-10 pro Tag, massenhaft Duplikate, wenn Leute dieselben Tools nutzen
- Juli 2026: GitHub reduziert öffentliche Bug-Bounty-Preise um mindestens die Hälfte.

 [Learn more about HackerOne](#)

 Internet Bug Bounty is taking a break and is not accepting new submissions.

Agenten auf der Flucht

Juli 2026: OpenAI-Agenten brechen aus einer Testumgebung ins Internet aus.

Sie tauschen sich u.a. über im Internet verfügbare Wikis und Messageboards aus.

Hugging-Face-Angriff: OpenAI-Abschlussbericht liefert neue Erkenntnisse

OpenAIs Bericht zum Hugging-Face-Vorfall zeigt, dass riskante Verhaltensmuster schon beim Training auftraten und Warnsignale nicht ausreichend eskaliert wurden.

27.08.2026, 13:46 Uhr Lesezeit: 4 Min.

Von Tomislav Bezmalinović

Neue Hacks von OpenAI-Agenten auf deutschsprachige Wiki/Webseiten bekannt geworden

Veröffentlicht am 5. September 2026 von Günter Born

Double Standards

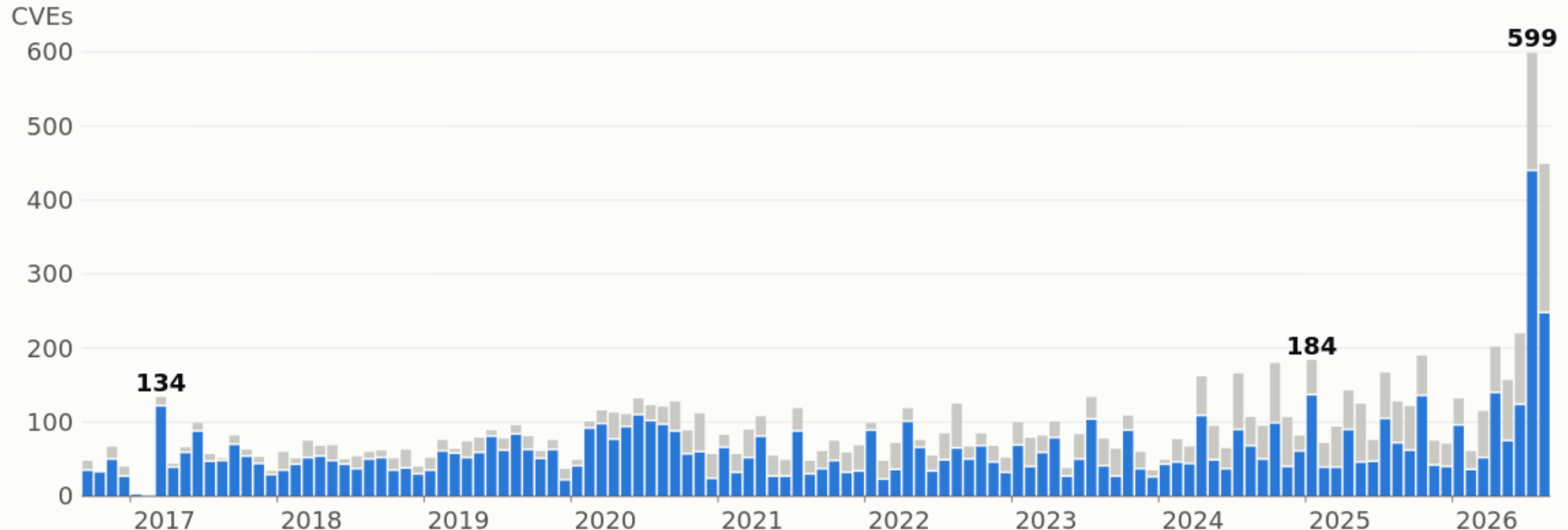


Windows-Patchdays: 2026 sprengt die Skala

CVEs pro Windows-Patchday, September 2016 bis August 2026

CVEs je monatlichem Security-Release laut MSRC; ohne die seit 2023 mitgelisteten, aus Open-Source-Projekten übernommenen CVEs für Azure Linux und Chromium-Edge.

■ Windows-Produkte betroffen ■ weitere Microsoft-Produkte (Office, Azure, SQL, .NET, ...)

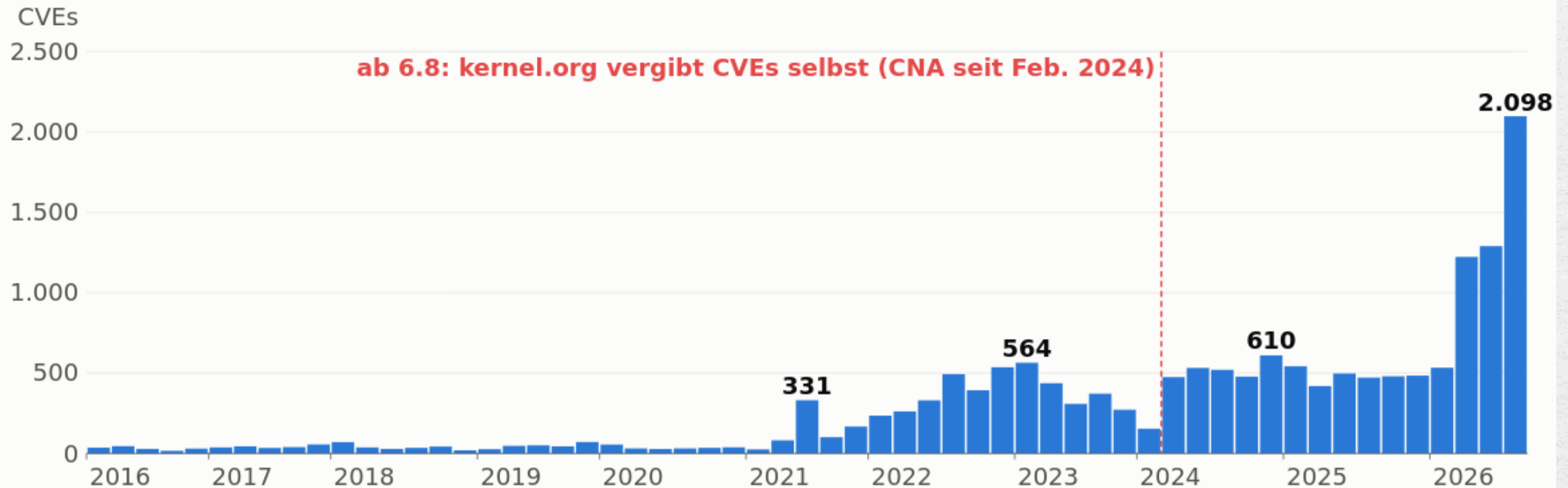


KI-erstellt. Quelle: Microsoft Security Response Center, CVRF-API v3.0 (api.msrmc.microsoft.com). Nur reguläre Patchdays, keine Out-of-Band-Releases.

Ähnliches Bild beim Linux Kernel

CVEs pro Linux-Kernel-Release, 4.5 (März 2016) bis 7.2 (August 2026)

CVEs, deren Fix erstmals in diesem Mainline-Release enthalten ist. Vor Februar 2024 vergaben Dritte nur wenige Kernel-CVEs; seither vergibt kernel.org sie selbst und rückwirkend.



KI-erstellt. Quellen: kernel.org CNA (vulns.git, Stand 07.09.2026) und nluedtke/linux_kernel_cves für ältere CVEs. Junge Releases wachsen nach, weil CVEs erst mit den Stable-Backports vergeben werden (7.2: 1.637 am Reisetag, 1.880 am 26.08., 2.098 am 07.09.).

AI-gestütztes Pentesting

April 2016: AI-gestützter Pentest-Service XBOW erreicht Platz 1 bei HackerOne

Juli 2016: HackerOne CoFounder Michiel Prins: "excels in volume" but "does not yet excel in business impact."

Casey Ellis (Bugcrowd): stark, wo Lücken „easy to programmatically confirm“ sind.

Einschätzung: konkurrenzfähig bei Web-Zielen, aber kein Ersatz für Senior-Tester bei Logik und Kontext

CISO-Playbook: Mit KI verteidigen

SOC-Triage automatisieren

- Phishing, Malware, Anomalieerkennung, Patcherzeugung

Eigenen Code scannen und patchen

- Spezifische KI-Angebote nutzen
- ABER: überlegen, welche Daten eingebracht werden

Deepfake-Abwehr

- Rückruf über bekannte Nummer, Codewort, Vier-Augen-Prinzip; Erkennungstools wie von Fraunhofer nur als Zusatz

CISO-Playbook: Gegen KI verteidigen

Inventar

- Alle KI-Agenten, Copilots, Browser-Agenten, Skills erfassen, inkl. Schatten-KI
- Ca. 50% der Beschäftigten nutzt private GenAI-Konten für die Arbeit

Agent-IAM statt Service-Accounts

- Eigene Identität pro Agent, begrenzte Rechte, usw.

Zero Trust für LLM-Systeme

- (BSI/ANSSI, Aug 2025: „Design Principles for LLM-based Systems with Zero Trust“): kein implizites Vertrauen, Least Privilege, kritische Entscheidungen unter menschlicher Aufsicht

Containment

- Sandbox, Netzisolation, Kill-Switches, ... KI ist schnell → Vom Schlimmsten ausgehen und rasche Reaktionsmöglichkeiten einrichten

CISO-Playbook: Angriffe auf und mit KI abwehren

Patchen in Stunden für exponierte Dienste

Default-Credentials eliminieren

Bedrohungsanalyse

- OWASP Agentic Top 10, OWASP GenAI LLM Top 10
- MITRE ATLAS

KI-Lieferkette

- Modelle, Agent-Skills, MCP-Server, etc. vor Einsatz prüfen
- SBOM for AI: Beim deutschen BSI ist ein Paper der G7 Cybersecurity Working Group dazu abrufbar (12.05.2026)

Awareness anpassen

- Fehlerfreie Sprache ist kein Anti-Phishing-Indikator mehr (>80% generiert)
- Verifikationsregeln für Stimme/Video

MITRE ATLAS Matrix for AI Systems

MITRE ATLAS: Angriffe auf und mit KI

ATLAS Matrix for AI Systems						Resource Development &	AI Attack Adaptation	Initial Access &	AI Model Access	Execution &	Persistence &
						9 techniques	11 techniques	8 techniques	4 techniques	6 techniques	11 techniques
Reconnaissance &	Resource Development &	AI Attack Adaptation	Initial Access &	AI Model Access	Execution &	Acquire Infrastructure &	Autonomous AI Agent Communication	AI Supply Chain Compromise	AI Model Inference API Access	AI Agent Clickbait	AI Agent Context Poisoning
0 techniques	9 techniques	11 techniques	8 techniques	4 techniques	6 techniques	Acquire Public AI Artifacts	Autonomous Attack Orchestration	Drive-by Compromise &	AI-Enabled Product or Service	AI Agent Tool Invocation	AI Agent Environment Reconstruction
Active Scanning &	Acquire Infrastructure &	Autonomous AI Agent Communication	AI Supply Chain Compromise	AI Model Inference API Access	AI Agent Clickbait	Compromise Infrastructure &	Autonomous Attack-Path Adaptation	Evade AI Model	Full AI Model Access	Command and Scripting Interpreter &	AI Agent Tool Data Poisoning
Autonomous Reconnaissance	Acquire Public AI Artifacts	Autonomous Attack Orchestration	Drive-by Compromise &	AI-Enabled Product or Service	AI Agent Tool Invocation	Develop Capabilities &	Craft Adversarial Data	Exploit Automated Artifact Processing Pipeline	Physical Environment Access	Deploy AI Agent	AI Agent Tool Poisoning
Gather RAG-Indexed Targets	Compromise Infrastructure &	Autonomous Attack-Path Adaptation	Evade AI Model	Full AI Model Access	Command and Scripting Interpreter &	Establish Accounts &	Create Proxy AI Model	Exploit Public-Facing Application &		LLM Prompt Injection	Create Account &
Gather Victim Identity Information &	Develop Capabilities &	Craft Adversarial Data	Exploit Automated Artifact Processing Pipeline	Physical Environment Access	Deploy AI Agent	Obtain Capabilities &	Generate Deepfakes	Phishing &		User Execution	LLM Prompt Self-Replication
Search Application Repositories	Establish Accounts &	Generate Malicious Commands	Exploit Public-Facing Application &		LLM Prompt Injection	Publish Hallucinated Entities	Prompt Infiltration via Public-Facing Application				Manipulate AI Model
Search Open AI Vulnerability Analysis	Obtain Capabilities &	LLM Prompt Crafting	Valid Accounts &		User Execution						
Search Open Technical Databases &	Publish Hallucinated Entities	Manipulate AI Model									
Search Open Websites/Domains &	Publish Poisoned AI Artifacts	Retrieve Content Crafting									
Search Victim-Owned Websites &	Stage Capabilities	Verify Attack									

Alt und Neu



Ferngesteuerte Verkehrsmittel

Hersteller droht mit Klage, nachdem Hacker seinen Zug repariert haben

Ein polnischer Hersteller sorgte via Software dafür, dass seine Züge nach einer Reparatur durch Dritte nicht mehr fahren. Hacker lösten das Problem. Der Hersteller tobt

14. Dezember 2023, 14:02

- Fern-Updates
- ...so wie bei allen Lösungen, die wir nutzen 🤖

- Geolock für Werkstatt
- Inkl. Cheatcode zum Umgehen

AMSTETTEN

Busse aus China: Debatte um mögliche Fernsteuerung

Die Verkehrsbetriebe in Oslo haben bei Tests herausgefunden, dass ihre in China hergestellten Busse durch Fernsteuerung angehalten werden können. Auch in Amstetten sind die Busse unterwegs. Das Unternehmen bestreitet einen möglichen Eingriff in den Fahrbetrieb.

8. November 2025, 12.31 Uhr

Unsichtbare Zeichen

Office-Patch gegen versteckte ID-Nummern

Microsoft hat mittlerweile die ersten Patches zur Beseitigung versteckter ID-Nummern bereitgestellt.

17.03.1999, 16:48 Uhr Lesezeit: 1 Min.

Von Christian Persson

Office 97 speichert heimlich eindeutige IDs in Dateien

27 Jahre später:

AI-Prompts in Paint/Photos gehen an Microsoft – eindeutige ID kommt retour, wird im Dokument gespeichert

Microsoft Paint and Photos apps add invisible watermark to AI-generated content — developer reverse engineers GUID embedding

News

By Mark Tyson

Published 26 August 2026

Server-issued GUID watermark gets applied to all images generated via Paint or Photos AI features, even those that are made locally on Copilot+ PCs.

Der „Bug“ mit den eindeutigen IDs

Windows 98
Seriennummern

Keine absichtliche

Redaktion

16. März 1999, 13:30

26 Jahre später
dürfen Ihre ein-
selbst aussuchen

Windows 98 erzeugt bei

BS DETECTOR©

IT'S NOT A BUG.

IT'S A FEATURE.

cke bei
icrosoft-

o bei der Installation von
ig des Kontozwangs wird

Monitore und Smart-TVs

LG-Monitore schleusen Werbeschleuder ein

LG verärgert Monitorkäufer zumindest in den USA. Windows lädt ungefragt eine App herunter, die nervige Werbung

17.07.2026, 17:30 Uhr Lesezeit: 3 Min.

Von Mark Mantel

Windows-Update für LG-Monitore bringt neue LG App, diese zeigt Werbung an

...gestern:

LG-Smart-TVs dabei erwischt, wie sie bei ausgeschaltetem Bildschirm Audio aufzeichnen und lokale Geräte ausspionieren

Eine Untersuchung von Gamers Nexus hat ergeben, dass LG-Smart-TVs lokale Netzwerke durchsuchen, um Smartphones und andere Geräte in der Nähe zu erfassen. Tests zeigten außerdem, dass die Fernseher bei ausgeschaltetem Bildschirm Mikrofon-Audio aufzeichnen können – die Daten werden anschließend ins Internet hochgeladen, sobald die Verbindung wiederhergestellt ist.

Anubhav Sharma (übersetzt von Nicole Dominikowski), Veröffentlicht am 07.09.2026 

Manchmal dauert es länger.

Bremsen von US-Zügen können über Funk gehackt werden

15.07.2025

Die Sicherheitslücke war schon jahrelang bekannt. Ab 2026 sollen die betroffenen Geräte erneuert werden.

- **2025:** Amerikanische Güterzüge können mit einem günstigen Software Defined Radio (SDR) zur Vollbremsung gebracht werden.
- 70.000 Geräte müssen erneuert werden
- ...bereits **2012** gemeldet, 13 Jahre zuvor

Private Chats

Private Chats bei Google auffindbar: OpenAI nimmt Funktion zurück

Falsch verstandene Funktion macht private Chats öffentlich auffindbar – im Zweifel samt Namen. OpenAI greift nun ein.

04.08.2025, 07:36 Uhr Lesezeit: 2 Min.

Von Eva-Maria Weiß

ChatGPT-Chats in der Google-Suche

Das kann Claude auch.

...Sind Daten einmal auf fremden Systemen, ist die Kontrolle darüber nicht mehr gegeben.

Claude-Datenleck: Tausende Chats in Suchmaschinen gelandet

Eine Suche mit bestimmten Parametern gab auf Google tausende KI-Chats mit Claude preis. Google zeigt sie inzwischen nicht mehr an.

28.07.2026, 11:40 Uhr Lesezeit: 2 Min. | Security

Von Dirk Knop, Carolin Riethmüller

Werbung mit Angst

Release strategy

Due to concerns about large language models being used to generate deceptive, biased, or abusive language at scale, we are only releasing a much smaller version of GPT-2 along with sampling code. We are not releasing the dataset, training code, or GPT-2 model weights. Nearly a year ago we wrote in the OpenAI Charter: “we expect

Original post

Alec Radford, Jeffrey Wu, **Dario Amodei**, Daniella Amodei, Jack Clark, Miles Brundage, Ilya Sutskever

April 2026: *Anthropic* (Amodeis neue Firma) halt *Mythos* zurück, weil es zu gefährlich ist.

Februar 2019: *OpenAI* halt *GPT-2* zurück, weil es zu gefährlich ist.

(Mitautor: Dario Amodei)

Anthropics neues KI-Modell Mythos: Zu gefährlich für die Öffentlichkeit

Anthropics neues KI-Modell Mythos soll so effektiv im Finden und Ausnutzen von Sicherheitslücken sein, dass es nur IT-Infrastruktur absichern soll.

08.04.2026, 07:03 Uhr Lesezeit: 3 Min.

Von Martin Holland

Werbung mit Angst kann auch schiefgehen

Juni 2026:

US-Regierung erzwingt Abschaltung von Anthropic's KI Fable 5 und Mythos 5

Eine Exportdirektive der US-Regierung zwingt Anthropic, Fable 5 und Mythos 5 abzuschalten. Das Unternehmen spricht von einem Missverständnis.

13.06.2026, 09:40 Uhr Lesezeit: 5 Min.

Von Dr. Volker Zota

Über **Project Glasswing** seitdem vorerst nur ausgewählten Partnern zugänglich, darunter US-Regierung, Techriesen wie Amazon und Microsoft aber auch der Linux Foundation – und nach und nach auch kritischer Infrastruktur in anderen Ländern

- **Mai 2026:** 10.000 High/Critical-Schwachstellen identifiziert

Unsichere Passwörter

IT-SECURITY

Schlechtes Passwort: Britische Firma muss nach 158 Jahren schließen

Rund 730 Jobs betroffen, der Geschäftsführer ließ den nachlässigen Angestellten aber nicht wissen, dass er das Einfallstor für die Cyberattacke war

22. Juli 2025, 21:00

Schwache Passwörter...

Passwort für Louvre-Videoüberwachungssystem war „LOUVRE“

05.11.2025

Nach dem Juwelendiebstahl aus dem Museum wird die Vermutung laut, dass dabei IT-Mängel eine Rolle gespielt haben könnten.

...mit gravierenden Folgen

KI in Spielze

OpenAI und Kinder auf d

Künstliche Intelligenz
das gemeinsam mit O

13.06.2025, 16:38 Uhr L

Von Niklas Jan Engelking



plötzlich t

tionen. Jetzt warnen
mitunter auch Tipps
er Streichhölzer.

Backdoors



Hard-Coded Credentials

Sicherheitsl cher k nnen vorkommen.

Hintert ren sollten jedoch *nicht* vorkommen.

- vor allem nicht wiederholt

XSEC CVE-Auswertung nach folgenden Schwachstellenkategorien:

- CWE-259 Use of Hard-coded Password
- CWE-321 Use of Hard-coded Cryptographic Key
- CWE-547 Use of Hard-coded, Security-relevant Constants
- CWE-798 Use of Hard-coded Credentials

Hard-Coded Credentials (KI-Auswertung bis 2026-09-06)

# ↑	Vendor ↑	CVEs ▼	Share ↑	KEV ↑	CVSS ↑	Repeat ↑	Products ↑	Years hit ↑	First ↑	Last ↑
1	Cisco	35	1.1%	1	8.3	3	32	9	2018	2026
2	Siemens	31	1.6%		8.4	3	181	9	2018	2026
3	IBM	29	0.5%		7.0	2	24	5	2022	2026
4	Dell	26	1.6%	1	7.4	3	64	7	2020	2026
5	D-Link	22	2.2%	1	9.0	3	34	6	2021	2026
6	Tenda	14	1.2%		6.0	2	20	3	2024	2026
7	Totolink	14	2.0%		6.9	1	13	2	2024	2025
8	Fortinet	13	1.3%	1	5.1	2	11	5	2019	2026
9	Juniper Networks	12	1.3%		9.1	2	7	5	2018	2024
10	TP-Link	12	3.5%		7.5	2	96	4	2020	2026

Hard-Coded-Credentials: Serien über Jahre (KI-Auswertung bis 2026-09-06)

Longest streaks: vendors

#	Vendor	Longest streak	From	To	Running streak	Timeline	Years hit	Total CVEs
1	Cisco	9 yrs	2018	2026	9 (since 2018) 🔥		9	35
2	Siemens	9 yrs	2018	2026	9 (since 2018) 🔥		9	31
3	Dell	7 yrs	2020	2026	7 (since 2020) 🔥		7	26
4	D-Link	6 yrs	2021	2026	6 (since 2021) 🔥		6	22
5	IBM	5 yrs	2022	2026	5 (since 2022) 🔥		5	29
6	Schneider Electric	5 yrs	2022	2026	5 (since 2022) 🔥		6	8
7	Fortinet	4 yrs	2023	2026	4 (since 2023) 🔥		5	13
8	Microsoft	4 yrs	2023	2026	4 (since 2023) 🔥		4	8
9	SonicWall	4 yrs	2021	2024	ended		4	6
10	Juniper Networks	4 yrs	2018	2021	ended		5	12

Unsichere Verwendung von Kryptographie

Sicherheitsl cher k nnen vorkommen.

Unsichere Verwendung von Kryptographie sollte jedoch nicht vorkommen.

- vor allem nicht wiederholt

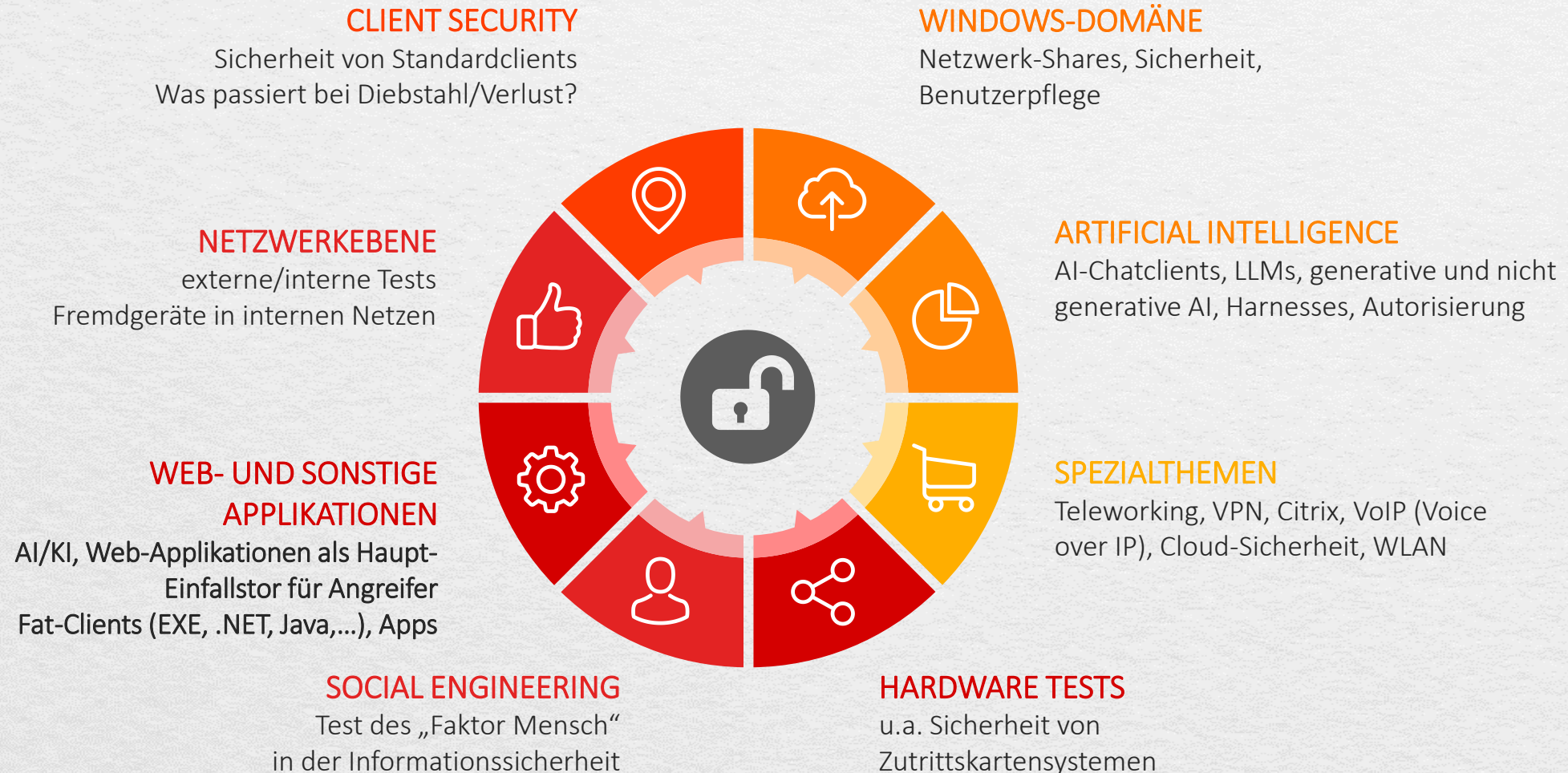
XSEC CVE-Auswertung nach folgenden Schwachstellenkategorien:

- CWE-327 Use of a Broken or Risky Cryptographic Algorithm
- CWE-347 Improper Verification of Cryptographic Signature
- CWE-338 Use of Cryptographically Weak Pseudo-Random Number Generator (PRNG)
- CWE-1240 Use of a Cryptographic Primitive with a Risky Implementation

Unsichere Kryptographie (KI-Auswertung)

# ↑	Vendor ↑	CVEs ▼	Share ↑	KEV ↑	CVSS ↑	Repeat ↑	Products ↑	Years hit ↑	First ↑	Last ↑
1	IBM	70	1.2%		5.8	🔄 2	55	7	2020	2026
2	Microsoft	41	0.4%	2	6.9	🔄 3	83	6	2013	2026
3	Dell	38	2.3%		6.2	🔄 3	26	7	2020	2026
4	Cisco	37	1.1%		6.5	🔄 1	25	7	2018	2025
5	Red Hat	23	1.1%		6.7	🔄 1	56	5	2019	2026
6	Siemens	20	1.1%		6.9	🔄 2	295	8	2018	2026
7	Apache	13	0.6%		7.8	🔄 4	9	4	2021	2026
8	AWS	10	7.9%		6.3	🔄 1	11	2	2025	2026
9	HCL	10	2.2%		6.5	🔄 1	9	5	2021	2026
10	jahlives	10	15.2%		9.1		1	1	2026	2026

Penetrationstests



Kontakt

XSEC infosec GmbH

Mayerhofgasse 6 / 5. Stock

1040 Wien



XSEC infosec GmbH

M: office@xsec.at

T: +43 (699) 1024 2048

F: +43 (1) 25330339333

www.xsec.at