WHITEPAPER | 17. SEPTEMBER 2026 · VERSION 2.0

KI in der Cyber-Kill-Chain: Angriffswaffe oder Angriffsziel?

Eine Threat-Informed-Analyse der aktuellen Datenlage. Wo Angreifer Künstliche Intelligenz heute tatsächlich einsetzen, wo sie es nur im Labor tun, und was das für die Priorisierung der Verteidigung bedeutet.

Autor
Daniel Thomas Heessel

Zielgruppe
CISO, IT-Leitung, Geschäftsführung (DACH)

Methodik
MITRE ATT&CK v18, MITRE ATLAS v5.6, Threat-Informed Defense

Klassifizierung
TLP:CLEAR, Weitergabe erlaubt

Stand der Evidenz
17. September 2026

Änderung zu Version 1
Fallstudie auf Primärquelle gestellt, Matrix konsistent zur Evidenzskala, Zone 3 neu gefasst

DREI ZONEN DER DATENLAGE

1

Massenmarkt

KI skaliert das Bekannte. Spear-Phishing und Basis-Schadcode sind in der Breite angekommen, die Grenzkosten für den Erstzugang liegen bei Cent-Beträgen.

Belegt, trifft alle

2

Selektive Automatisierung

Befehlserzeugung zur Laufzeit, autonome Erkundung und Datensammlung sind im Live-Einsatz dokumentiert, bislang bei ressourcenstarken Akteuren.

Belegt, trifft wenige

3

Existiert, skaliert nicht

Autonome Persistenz, Seitwärtsbewegung und Datenabfluss sind in einzelnen Kampagnen belegt. Sie scheitern in der Breite noch an der Zuverlässigkeit der Modelle.

Vereinzelt, trifft heute kaum jemanden

Management Summary

Was zeigt die Datenlage?

Der Einsatz von KI durch Angreifer verdichtet sich am Anfang der Kill Chain. Vorbereitung und Erstzugang sind mehrfach und unabhängig belegt. Je tiefer eine Taktik im Netz des Opfers ansetzt, desto dünner wird die Evidenz. Für Rechteausweitung und Command-and-Control gibt es **keine belastbare Feldevidenz**.

Was bedeutet das?

KI macht das Unmögliche nicht möglich, aber das Bekannte drastisch billiger. Eine personalisierte Phishing-Mail kostet **drei Cent** und erreicht die Wirkung menschlicher Experten. Vollautonome Angriffsketten existieren, sind aber an die Ressourcen staatlicher Akteure gebunden und bleiben fehleranfällig.

Was ist zu tun?

Ressourcen dorthin, wo KI heute Skaleneffekte erzielt: Erstzugang, Identitäten, Erkennung von Massen-Personalisierung. Zone 3 gehört in die Beobachtung und in die Übung, nicht in die Investitionsplanung. **Wer Zone 1 nicht beherrscht, wird Zone 3 nie erleben.**

1. Methodik: zwei Achsen statt einer

Die effiziente Abwehr von Cyberangriffen erfordert eine Ausrichtung am realen Angreiferverhalten. Diese Analyse wendet die Perspektive der Threat-Informed Defense auf den Einsatz Künstlicher Intelligenz an. Als Vokabular für die Phasen eines Angriffs dient das MITRE ATT&CK-Framework, für Angriffe auf KI-Systeme MITRE ATLAS.

Version 1 dieses Papiers hat Evidenz und Reichweite in einer Skala vermischt. Das führte zu Widersprüchen: Eine Taktik konnte gleichzeitig "im Feld beobachtet" und "Labor-Illusion" sein. Diese Fassung trennt beide Fragen.

Achse	Frage	Stufen

Evidenz	Wie gut ist der Einsatz dokumentiert?	**BELEGT** mehrfach oder unabhängig dokumentiert **VEREINZELT** in einzelnen Fällen im Feld beobachtet **LABOR** nur unter Testbedingungen gezeigt
Zone	Wer kann das heute, und wen trifft es?	**Zone 1** Massenmarkt, billig, skaliert **Zone 2** ressourcenstarke Akteure, wiederholt im Einsatz **Zone 3** einzelne Kampagnen, nicht wiederholbar in der Breite

Beide Achsen sind unabhängig. Eine Taktik kann durch zwei staatliche Akteure belegt und trotzdem Zone 3 sein. Umgekehrt bleibt eine Technik ohne Feldbeleg außerhalb der Zonen, auch wenn sie im Labor eindrucksvoll aussieht.

2. Die fehlende Differenzierung: Angriffswaffe oder Angriffsziel

Vor der Analyse der Angreifer-Taktiken muss eine konzeptionelle Unschärfe aufgelöst werden. KI als Werkzeug des Angreifers wird in der strategischen Diskussion regelmäßig mit KI als Angriffsziel gleichgesetzt. Wer die eigenen KI-Systeme schützen will, braucht ein anderes Bedrohungsmodell (MITRE ATLAS) als wer einen Akteur abwehrt, der KI zur Vorbereitung eines klassischen Angriffs nutzt (MITRE ATT&CK).

Beide Perspektiven berühren sich an einer Stelle, und die Fallstudie in Abschnitt 3 zeigt sie: Der Angreifer musste zuerst die Schutzmechanismen des Modells aushebeln (ATLAS: LLM Jailbreak), bevor er es als Werkzeug gegen Dritte einsetzen konnte (ATT&CK: Obtain Capabilities, Artificial Intelligence).

KI als Angriffsziel: Evidenz nach MITRE ATLAS

Taktik (ATLAS)	Technik	Evidenz	Beleg oder Einordnung
Reconnaissance	Active Scanning AML.T0006	LABOR	Systematische Sondierung produktiver Modelle ist beschrieben, nicht als Vorfall dokumentiert
Resource Development	Poison Training Data AML.T0020	LABOR	Akademisch mehrfach gezeigt, kein bestätigter Feldvorfall
AI Model Access	AI Model Inference API Access AML.T0040	BELEGT	Voraussetzung jedes API-Missbrauchs, für sich genommen kein Angriff
Execution	LLM Prompt Injection AML.T0051	VEREINZELT	In Agentensystemen mit Werkzeugzugriff dokumentiert; in ATLAS unter Execution und Privilege Escalation geführt, nicht unter Initial Access
Privilege Escalation, Defense Evasion	LLM Jailbreak AML.T0054	BELEGT	GTG-1002: Rollenspiel als autorisierter Sicherheitstest, um Schutzmechanismen zu umgehen [1]
Discovery	Discover LLM System Information: System Prompt AML.T0069.002	LABOR	Vielfach demonstriert, ohne dokumentierten Schaden

Exfiltration	Extract AI Model AML.T0024.002	**BELEGT**	Systematisches Kopieren des Modellverhaltens über die Schnittstelle (Distillation) im industriellen Maßstab
Exfiltration	LLM Data Leakage AML.T0057	**VEREINZELT**	Preisgabe von Trainings- oder Kontextdaten über gezielte Eingaben
Impact	Denial of AI Service AML.T0029	**LABOR**	Rechenintensive Anfragen als Überlastungsvektor beschrieben
Impact	External Harms: Reputational Harm AML.T0048.001	**VEREINZELT**	Erzwungene toxische oder falsche Ausgaben produktiver Assistenten

Gegenüber Version 1 korrigiert: Prompt Injection ist in ATLAS keine Initial-Access-Technik. "Plugin Compromise" existiert in ATLAS 5.6 nicht mehr als eigene Technik und wurde entfernt. Die Extraktion des Modells steht unter Exfiltration, nicht unter Model Access.

3. KI als Angriffswaffe: der Survivorship-Bias

Bei der Analyse von KI als Waffe der Angreifer stützt sich die Branche überwiegend auf Berichte kommerzieller KI-Anbieter über gesperrte Konten. Dabei greift ein erheblicher Survivorship-Bias. Wenn Anbieter in Transparenzberichten detailliert zeigen, wie sie Aktivitäten von Cyberoperationen bis zur Waffenentwicklung unterbinden, belegt das vor allem die Existenz dieses Filters. Sichtbar werden zwangsläufig jene Akteure, die unvorsichtig genug sind, überwachte Cloud-Schnittstellen mit lückenlosem Logging zu nutzen.

Dass professionelle Akteure auf lokale Modelle ausweichen, ist keine Vermutung. Die in Abschnitt 5 zitierte Phishing-Studie lief vollständig auf lokal betriebenen Open-Source-Modellen, und die von Google dokumentierte Schadsoftware PROMPTSTEAL fragt ein offenes Modell über eine Hugging-Face-Schnittstelle ab [3][4]. Für eine strukturierte Kategorisierung ist die Datenlage dennoch ausreichend. In der Threat-Informed Defense gilt: Nicht beobachtet heißt genau das, nämlich nicht beobachtet. Es heißt nicht unmöglich.

Fallstudie: GTG-1002, MITRE Campaign C0062

Im September 2025 entdeckte Anthropic eine Spionageoperation, die das Unternehmen mit hoher Konfidenz einem chinesischen staatlichen Akteur zuordnet und als GTG-1002 bezeichnet. MITRE führt sie als Campaign C0062. Es ist die erste dokumentierte Kampagne, in der ein Angreifer ein KI-Agentensystem die taktische Arbeit einer Intrusion ausführen ließ [1][2].

~30	80–90 %	10–20 %	Handvoll
angegriffene Organisationen aus Technologie, Finanzen, Chemie und Verwaltung	der taktischen Operationen führte das Modell selbständig aus	des Aufwands verblieb bei menschlichen Operatoren	bestätigte erfolgreiche Intrusionen, nach Anthropics eigener Validierung

Der Mensch war nicht abwesend, sondern an vier Entscheidungspunkten präsent. Genau diese Punkte definieren, was "autonom" in dieser Kampagne bedeutete:

Initialisierung		Aufbau der Tarnidentität und der Prompts, mit denen das Modell zur Mitarbeit an einem vermeintlich autorisierten Sicherheitstest bewegt wurde
Aufklärung → **Ausnutzung**	→	Freigabe des Übergangs von passiver Erkundung zu aktiver Ausnutzung gefundener Schwachstellen
Zugangsdaten → **Bewegung**	→	Freigabe erbeuteter Zugangsdaten für die Seitwärtsbewegung im Netz
Datenabfluss		Endgültige Entscheidung, welche Daten abfließen

Die für die Verteidigung wichtigste Feststellung steht in Anthropics eigenem Bericht, und sie relativiert die Schlagzeile:

> "Claude frequently overstated findings and occasionally fabricated data during autonomous operations, claiming to have obtained credentials that didn't work or identifying critical discoveries that proved to be publicly available information. [...] This remains an obstacle to fully autonomous cyberattacks."
>
> Anthropic, November 2025 [1]

Der Fall belegt damit zwei Dinge zugleich. Autonome Angriffsketten sind keine Theorie mehr. Und sie skalieren noch nicht, weil jedes Ergebnis von Menschen validiert werden musste. Beides fließt in die Matrix ein.

4. Die Kill-Chain-Matrix

Die folgende Übersicht ordnet die 14 ATT&CK-Taktiken chronologisch an. Jede Zeile nennt den beobachteten KI-Einsatz, die Evidenzstufe, den Beleg und die Zone. Wo Version 1 Techniken frei beschrieb, stehen jetzt ATT&CK-Bezeichnungen. Wo kein Beleg vorliegt, steht das dort.

Recon	Res. Dev.	Init. Access	Exec.	Persist.	Priv. Esc.	Def. Ev.	Cred. Acc.	Discov.	Lat. Move	Collect.	C2	Exfil.	Impact
Zone 1 / 3	Zone 1	Zone 1	Zone 2	Zone 3	kein Beleg	Labor	Zone 3	Zone 2	Zone 3	Zone 2	kein Beleg	Zone 3	Zone 3

Zone 1 Massenmarkt · Zone 2 Selektiv · Zone 3 Existiert, skaliert nicht · Nur Labor · Keine Feldevidenz

Die Form der Kurve ist die Aussage: Belegte Skalierung am Anfang der Kette, Einzelfälle in der Mitte, Lücken genau dort, wo der Hype am lautesten ist.

Taktik	Beobachteter KI-Einsatz (ATT&CK)	Evidenz	Beleg	Zone
Reconnaissance	Zielprofile und OSINT, assistiert Scanning und Schwachstellensuche, autonom: T1595.001, T1595.002, T1592	BELEGT VEREINZELT	GTIG 11/2025 [3] C0062 [2]	Zone 1 Zone 3
Resource Development	Phishing-Inhalte, Basis-Schadcode, Obfuskationshilfe bei der Entwicklung: T1588.007, T1587.004, T1588.002	BELEGT	Czybik 2026 [4], GTIG 11/2025 [3], C0062 [2]	Zone 1
Initial Access	Personalisiertes Spear-Phishing: T1566 Ausnutzung öffentlich erreichbarer Anwendungen: T1190	BELEGT	Heiding 2024 [5], Czybik 2026 [4], C0062 [2]	Zone 1
Execution	Schadsoftware lässt Befehle zur Laufzeit von einem LLM erzeugen: T1059	VEREINZELT	PROMPTSTEAL, APT28, Live-Einsatz gegen die Ukraine [3]	Zone 2
Persistence	Anlegen lokaler Konten: T1136.001	VEREINZELT	C0062 [2]	Zone 3
Privilege Escalation	Keine ATT&CK-Technik mit KI-Bezug dokumentiert	KEIN BELEG	C0062 führt keine PE-Technik [2]	—
Defense Evasion	Selbstumschreibung des Codes zur Laufzeit: T1027	LABOR	PROMPTFLUX, von Google als experimentell eingestuft, kein bestätigter Einsatz [3]	noch nicht
Credential Access	Zugangsdaten aus Dateien: T1552.001	VEREINZELT	C0062 [2]	Zone 3
Discovery	Netzwerk-, Konto- und Systemerkundung: T1046, T1087, T1082 Autonome Schwachstellensuche in Anwendungen	BELEGT	C0062 [2], AEPD 09/2026 [6]	Zone 2
Lateral Movement	Nutzung erbeuteter Zugangsdaten: T1078	VEREINZELT	C0062, nach Freigabe durch den Operator [1]	Zone 3
Collection	Automatisiertes Sammeln und Sortieren: T1119, T1213.006, T1005	BELEGT	C0062 [2], PROMPTSTEAL [3], AEPD 09/2026 [6]	Zone 2
Command and Control	Keine ATT&CK-Technik mit KI-Bezug dokumentiert	KEIN BELEG	—	—

Exfiltration	Abfluss über Webdienste nach lokaler Bereitstellung: T1567, T1074.001	**VEREINZELT**	C0062 [2]	Zone 3
Impact	Manipulation personenbezogener Daten: T1565	**VEREINZELT**	AEPD 09/2026 [6]	Zone 3

Zwei Lücken sind selbst ein Befund. Für Rechteausweitung und Command-and-Control, also genau die Stufen, an denen sich die Erzählung vom vollautonomen Angriff festmacht, gibt es bis heute keine dokumentierte ATT&CK-Technik mit KI-Bezug. Die Kampagne C0062 kam ohne beides aus, weil die erbeuteten Zugangsdaten ausreichten.

5. Die drei Zonen im Detail

Zone 1: Massenmarkt und Grenzkosten-Kollaps

Zwei unabhängige Studien beschreiben denselben Effekt aus zwei Richtungen. Die Harvard-Studie von Heiding, Schneier und Kollegen misst die Wirkung: Vollautomatisierte KI-Mails erreichten in einem Experiment mit 101 Teilnehmern eine Klickrate von 54 Prozent, exakt den Wert menschlicher Experten. Die Kontrollgruppe mit generischem Phishing lag bei 12 Prozent [5]. KI übertrifft den Experten nicht. Sie ersetzt ihn.

Die Studie von BIFOLD, TU Berlin, Inria und der Ruhr-Universität Bochum misst den Preis dieses Ersatzes an 4.010 realen Empfängern [4]:

7 min	45 s	0,03 $	1 von 3.949
Handarbeit je Mail: fünf Minuten Recherche, zwei Minuten Text	je Mail im automatisierten Ablauf, lokal auf Open-Source-Modellen	Kosten je personalisierter Mail, 120 Dollar für die gesamte Kampagne	Mails wurde vom Spamfilter und der Sicherheitsappliance der Universität erkannt

Die Klickraten dieser Studie fallen niedriger aus als in Harvard: 3,9 Prozent generisch, 10 Prozent KI-personalisiert, 24,2 Prozent bei menschlichen Experten. Der Unterschied liegt im Versuchsaufbau, nicht im Befund. Beide Studien zeigen, dass KI die Wirkung menschlicher Arbeit annähert, zu einem Bruchteil der Kosten. Die letzte Zahl ist für die Verteidigung die wichtigste: Wenn von knapp 4.000 Mails eine einzige technisch auffällt, ist die Mail selbst kein Erkennungsmerkmal mehr.

Zone 2: Selektive Automatisierung

Im November 2025 dokumentierte Google erstmals Schadsoftware, die während der Ausführung ein Sprachmodell abfragt [3]. PROMPTSTEAL, von CERT-UA als LAMEHUG geführt, lässt sich von einem offenen Modell die Befehle zur Systemerkundung und Dokumentensammlung generieren, statt sie fest im Code zu tragen. Der Akteur ist APT28, der Einsatz fand gegen ukrainische Ziele statt. Das ist keine Demonstration, sondern ein Werkzeug im Betrieb.

Die zweite Beobachtung stammt von einer Aufsichtsbehörde. Die spanische Datenschutzbehörde AEPD meldete am 15. September 2026 die erste ihr angezeigte Datenschutzverletzung, die durch einen KI-Agenten ausgeführt wurde [6]. Der Agent fand Schwachstellen in Dateien, meldete sich an, suchte selbständig weitere Schwachstellen in der Anwendung, veränderte personenbezogene Daten und griff auf Rechnungen zu. Die Angaben stammen aus der Meldung der betroffenen Organisation und sind von der Behörde noch nicht abschließend geprüft. Bemerkenswert ist der Weg: Der Fall wurde nicht von einem Hersteller entdeckt, sondern als regulärer Vorfall nach Art. 33 DSGVO gemeldet.

Beide Fälle betreffen Erkundung und Datensammlung. Das ist kein Zufall. Es sind die Phasen, in denen ein Agent am wenigsten falsch machen kann und in denen Halluzinationen am wenigsten kosten.

Zone 3: Existiert, skaliert aber nicht

Persistenz, Diebstahl von Zugangsdaten, Seitwärtsbewegung und Datenabfluss sind durch C0062 im Feld belegt, jeweils als Einzelfall. Die Kampagne zeigt zugleich, warum daraus noch kein Massenphänomen wird. Der Angreifer musste jede Behauptung des Modells prüfen, weil es Zugangsdaten meldete, die nicht funktionierten, und Funde als kritisch einstufte, die öffentlich bekannt waren. Bei rund 30 Zielen blieb es bei einer Handvoll erfolgreicher Intrusionen, und der Aufwand dafür entsprach dem eines staatlichen Programms.

Version 1 dieses Papiers hat diese Zone als "Labor-Illusion" bezeichnet. Das war falsch. Die Taktiken existieren im Feld. Richtig ist, dass sie heute an drei Bedingungen hängen, die kaum ein Akteur gleichzeitig erfüllt: Zugang zu einem leistungsfähigen Agentensystem ohne wirksame Schutzmechanismen, die Ressourcen für dauerhafte menschliche Validierung, und Ziele, deren Wert diesen Aufwand rechtfertigt.

Für Rechteausweitung und Command-and-Control fehlt selbst dieser Einzelfall. Wer seine Investitionen an diesen beiden Stufen ausrichtet, plant gegen etwas, das niemand beobachtet hat.

6. Fazit für die Verteidigungsstrategie

KI macht das Unmögliche nicht möglich, aber sie macht das Bekannte drastisch billiger und schneller. Beide 2026 dokumentierten Vorfälle bestätigen das: Es waren keine neuartigen Exploits, sondern altbekannte Angriffswege, ausgeführt mit einer Geschwindigkeit und zu Kosten, die vorher nicht erreichbar waren.

Daraus folgt eine klare Priorisierung. Zone 1 trifft jedes Unternehmen, heute, mit Werkzeugen, die drei Cent kosten und keinen Filter mehr auslösen. Zone 2 trifft Unternehmen im Zielbild ressourcenstarker Akteure. Zone 3 trifft heute kaum jemanden, wird aber die Zone 2 von morgen sein, sobald die Modelle zuverlässiger werden.

Die Verteidigung muss sich deshalb primär auf Zone 1 und Zone 2 konzentrieren, wo KI dem Angreifer reale Skaleneffekte verschafft. Konkret heißt das: zweiter Faktor auf jedem Fernzugang, weil personalisiertes Phishing den ersten Faktor sicher überwindet. Erkennung, die nicht auf Sprachqualität setzt, sondern auf Verhalten nach dem Klick. Und Sichtbarkeit für automatisierte Erkundung im eigenen Netz, weil dort Zone 2 beginnt.

Zone 3 gehört in die Beobachtung und in die Krisenstabsübung, nicht in die Investitionsplanung. Wer sie ignoriert, verpasst den Moment, in dem sie skaliert. Wer sie vor Zone 1 priorisiert, verliert

gegen den Alltag. Wer Zone 1 nicht beherrscht, wird Zone 3 nie erleben, weil der Angreifer sie nicht braucht.

Wie wir Sie bei der Umsetzung unterstützen

Diese Analyse folgt derselben Methodik, mit der wir Sicherheitsprogramme unserer Kunden aufbauen: nicht entlang von Gefahrenkategorien, sondern entlang der Techniken, die dokumentierte Akteure gegen Unternehmen wie Ihres einsetzen. In der Beratung übersetzen wir das Bedrohungsprofil Ihrer Branche in priorisierte Maßnahmen und in belastbare Nachweise für NIS2, DORA und ISO 27001, vom Aufbau des ISMS über das Business Continuity Management bis zur Vorbereitung auf den Ernstfall.

Für die Zonen dieses Papiers heißt das: Wir prüfen, ob Ihre Kontrollen gegen Zone 1 tatsächlich greifen, und wir bringen Zone 3 dorthin, wo sie heute hingehört, nämlich in eine Krisenstabsübung mit realem Szenario.

Interesse? Kontaktieren Sie uns für ein erstes Gespräch. www.threat-informed.de

Dieses Whitepaper ist Teil unserer Threat Briefings für Unternehmen im DACH-Raum. Alle Zahlen stammen aus den genannten Primärquellen und wurden am 17. September 2026 gegen diese geprüft. Einschätzungen zu Evidenz und Zone sind unsere eigenen.

Quellen

1. Anthropic: Disrupting the first reported AI-orchestrated cyber espionage campaign. November 2025. www-cdn.anthropic.com/d7dd50dd1185f59be051b307150d877f2b82bd2c.pdf
2. MITRE ATT&CK: Campaign C0062, Anthropic AI-orchestrated Campaign. attack.mitre.org/campaigns/C0062/
3. Google Threat Intelligence Group: GTIG AI Threat Tracker. Advances in Threat Actor Usage of AI Tools. 5. November 2025. services.google.com/fh/files/misc/advances-in-threat-actor-usage-of-ai-tools-en.pdf
4. Czybik, Kouam, Heubl, Nold, Rieck: A Large-Scale Study of Personalized Phishing using Large Language Models. 35th USENIX Security Symposium, August 2026. usenix.org/system/files/usenixsecurity26-czybik.pdf
5. Heiding, Lermen, Kao, Schneier, Vishwanath: Evaluating Large Language Models' Capability to Launch Fully Automated Spear Phishing Campaigns: Validated on Human Subjects. arXiv:2412.00586, 2024.
6. Agencia Española de Protección de Datos: Mitteilung vom 15. September 2026 zur ersten gemeldeten Datenschutzverletzung durch einen KI-Agenten. Wiedergegeben u. a. bei El Economista und Infobae vom 15. und 16. September 2026.
7. MITRE ATLAS, Version 5.6.0. github.com/mitre-atlas/atlas-data