REPORT

KI-Sicherheit 2026

Erkenntnisse aus über 625.000 Kundenumgebungen weltweit

SOPHOS

Vorwort von John Peterson

Im Mai 2026 entdeckte Sophos einen Angreifer, der im Netzwerk eines Kunden quasi eine Softwareentwicklungsabteilung betrieb. Etwa ein Dutzend KI-Agenten, die über einen kommerziellen Programmierassistenten koordiniert wurden, schrieben und testeten Malware gegen den Endpoint-Schutz von Sophos, CrowdStrike und Windows Defender, jeweils auf einer eigenen virtuellen Maschine. Auf diese Weise entstanden fast 80 Module und mehr als 70 Evasionstechniken, wobei jedes Ergebnis in die Versionskontrolle übernommen und beim nächsten Durchgang verbessert wurde. Derselbe Betreiber nutzte die Ergebnisse später, um Ransomware bereitzustellen und Daten zu stehlen.

Zwar war die Arbeitsweise an sich nicht neu, doch gelang es den Agenten, wochenlange manuelle Iterationen auf wenige Tage automatisierter Iterationen zu verkürzen – genau diese Veränderung ist Gegenstand dieses Reports. KI verändert die Geschwindigkeit und die Ausgestaltung von Security Operations stärker als die grundlegenden Mechanismen von Angriffen. Angreifer benötigen nach wie vor einen Erstzugriff, bewegen sich weiterhin lateral und exfiltrieren Daten nach wie vor über beobachtbare Kanäle. Was sich geändert hat, ist die Zeit.

In diesem Report wird Sophos dies anhand von Daten belegen.

Unsere Analysten für Managed Detection and Response (MDR) und Incident Response (IR) analysieren jährlich Tausende von Vorfällen. Unsere Counter Threat Unit (CTU) überwacht Untergrundforen in mehreren Sprachen. Unsere SophosLabs-Forscher analysieren Malware-Samples in großem Maßstab. Unser Sophos AI Team entwickelt innovative, abwehrorientierte Techniken. Und wir verfügen über Endpoint-, Firewall-, -E-Mail-, Cloud-, Netzwerk- und Identitäts-Telemetriedaten aus über 625.000 Kundenumgebungen.

Diese Größe, Reichweite und Expertise verschaffen uns eine Perspektive, über die nur wenige andere an der sich rasch entwickelnden Schnittstelle zwischen KI und Cybersicherheit verfügen. Wir beabsichtigen, diese Perspektive zu nutzen, um der Branche zu helfen, der Zeit voraus zu sein und sicherzustellen, dass unsere Kunden geschützt bleiben.

J. P. Peterson, CTO, Sophos

Zusammenfassung und wichtige Erkenntnisse

Dieser Report stützt sich auf Sophos MDR- und IR-Fälle, SophosLabs-Analysen, Sophos CTU Intelligence sowie Endpoint- und Netzwerkbeobachtungen in über 625.000 Kundenumgebungen weltweit. Ein Muster zieht sich wie ein roter Faden durch die gesamte Entwicklung: KI ermöglicht es Angreifern und Verteidigern, vertraute Abläufe schneller zu durchlaufen. Allerdings gibt es bislang nur relativ wenig Hinweise darauf, dass mittels KI komplett neue Angriffsarten entwickelt werden. Mit zunehmender Leistungsfähigkeit von KI-Modellen könnte sich dies jedoch schon bald ändern.

2026 ist ein Wendepunkt. Frontier- und Open-Weight-Modelle haben sich für die Delegation von Entwicklungsaufgaben als ausreichend nützlich erwiesen. Gleichzeitig dazu verändern sinkende Inferenzkosten die wirtschaftlichen Rahmenbedingungen für beide Seiten. Die praktische Entscheidung ist nicht mehr, ob KI eingesetzt wird, sondern wo teure Frontier-Systeme gerechtfertigt sind, wo günstigere Open-Weight-Modelle ausreichen und wie Workflows ohne Risikoerweiterung gesteuert werden können.

Die folgenden Erkenntnisse bilden die Grundlage des Reports:

1. **KI macht Angriffe schneller, erfindet selbst aber keine neuen Angriffsarten.** Der von Sophos als STAC6994 bezeichnete Angreifer nutzte KI-Agenten, um EDR-Umgehungstechniken in einem Tempo weiterzuentwickeln, für das ein menschlicher Entwickler Wochen gebraucht hätte. Sophos CTU-Analysten bestätigten anschließend, dass der Betreiber hinter STAC6994 an der Bereitstellung von Ransomware und Datendiebstahloperationen beteiligt war.
2. **Zugangsdaten und Identitätssysteme rund um KI-Services in Unternehmen sind mittlerweile ein Ziel für Datendiebstahl geworden.** Der Salesloft/Drift-Vorfall hat gezeigt, wie OAuth-Token von Chatbots direkt zu einer Kompromittierung der Salesforce-Umgebung führen können. Unabhängig davon haben Sophos CTU-Analysten in Untergrundforen Behauptungen von Angreifern beobachtet, wonach KI-Prompts und generierte Daten bei Cyberangriffen als Kollateralschaden erbeutet werden könnten.
3. **Die Schattenwirtschaft integriert KI in ihre Infrastrukturen.** Sophos hat dokumentiert, dass Angreifer API-Schlüssel für Claude, ChatGPT und Grok über Zwischenhändler verkaufen. In Foren sind Kanäle entstanden, die sich mit KI-Prompt-Engineering, Jailbreaking-Techniken und Workflows zur Malware-Entwicklung befassen. Einige Akteure stellen KI-Prompt-Engineers ein. Andere bleiben offen skeptisch, dass KI ihre Arbeitsabläufe verändern wird. Die Adoptionskurve sieht weniger nach einer plötzlichen Transformation aus, sondern eher nach der schrittweisen Einbindung eines neuen Tools in bestehende kriminelle Workflows.
4. **Supply-Chain-Angriffe nehmen gezielt KI-Entwicklungsstrukturen ins Visier.** Die Kompromittierung der Nx Console VS Code-Erweiterung verbreitete Malware zum Diebstahl von Zugangsdaten. Analysten der SophosLabs analysierten gefälschte Claude-Download-Seiten, über die ein neuartiger RAT verbreitet wurde, sowie eine ClickFix-Kampagne, bei der legitime, über ChatGPT geteilte Chat-URLs genutzt wurden, um den Infostealer „MacSync“ zu verbreiten.
5. **KI-gestütztes Social Engineering kommt nicht mehr nur experimentell zum Einsatz, sondern wird standardmäßig angewendet.** Der Sophos CTU liegen Untergrundanzeigen für Sprachbots, KI-generierte Identitäten für Liebesbetrug sowie Services für synthetische Profilbilder vor. Der Beitrag der KI besteht in der sprachübergreifenden Skalierung und sprachlichen Qualität, angewendet auf dieselben Täuschungstechniken, die bereits vor der generativen KI funktionierten.

6. **Schwachstellen werden immer schneller ausgenutzt als sie gepatcht werden.** Die am 10. Juni 2026 herausgegebene Binding Operational Directive 26-04 der CISA kürzte die obligatorische Patch-Frist für die risikoreichsten Schwachstellen auf drei Tage und nannte explizit KI als einen Faktor, der das Zeitfenster zwischen Offenlegung und Ausnutzung verengt. Der erste Praxistest für die Richtlinie kam fast sofort: CVE-2026-10520 (Ivanti Sentry) wurde innerhalb von 24 Stunden nach der Proof-of-Concept-Veröffentlichung ausgenutzt und am 12. Juni in den KEV-Katalog aufgenommen. CVE-2026-42208 (LiteLLM) wurde innerhalb von 36 Stunden ausgenutzt, wobei Angreifer Datenbanken ins Visier nahmen, die Schlüssel von Upstream-LLM-Anbietern enthielten.

7. **Der Nachweis eines KI-Einsatzes bei einem Angriff bleibt operativ schwierig.** Der Fall STAC6994 war ungewöhnlich, da das Entwicklungs-Framework auf einem Gerät zurückgelassen wurde, das Sophos untersuchen konnte. Bei vielen Vorfällen ist KI-Unterstützung in der Telemetrie unsichtbar. Die tatsächliche Verbreitung von KI-gestützten Angriffen ist wahrscheinlich höher als jeder Anbieter direkt belegen kann.

8. **KI-Reasoning-Modelle beschleunigen die Analyse.** KI-Reasoning-Modelle beschleunigen die Analyse bei bekannten Angriffsmustern von Stunden auf Minuten. Sie sind kein Ersatz für Verhaltenserkennungssysteme, die eine einzelne Bedrohung in Milliarden von Ereignissen aufspüren. Die statistischen Einschränkungen aufgrund eines extremen Klassenungleichgewichts, die Notwendigkeit umgebungsspezifischer Basislinien und das Tempo der Anpassung durch Angreifer erfordern allesamt spezielle technische Lösungen

Teil 1:
KI in der Angriffskette

Sophos X-Ops-Taxonomie von KI-Bedrohungen

Sophos X-Ops unterteilt KI-Bedrohungen in zwei übergeordnete Kategorien: missbräuchliche Nutzung von KI (die Bedrohungen, vor denen wir schützen) und gezielte Angriffe auf KI (die KI, die wir schützen). Die erste umfasst KI-generierte Artefakte, KI-erweiterte Laufzeitfunktionen und KI-orchestrierte Eindringversuche. Die zweite umfasst von Agenten initiierte Kompromittierungen, Impersonation von KI-Software, manipulierte LLMs und Angriffe auf die Modelle selbst. Für beide Kategorien sind jeweils unterschiedliche Gegenmaßnahmen erforderlich: Missbräuchliche Nutzung ist ein Problem der Produktivität und der Skalierbarkeit; gezielte Angriffe sind ein Problem des Vertrauens, der Lieferkette und der Governance. Die Taxonomie ergänzt bestehende Frameworks wie MITRE ATLAS, Microsofts Taxonomie für Agenten-Fehlermodi, MITs AI Risk Repository und NIST AI 100-2.

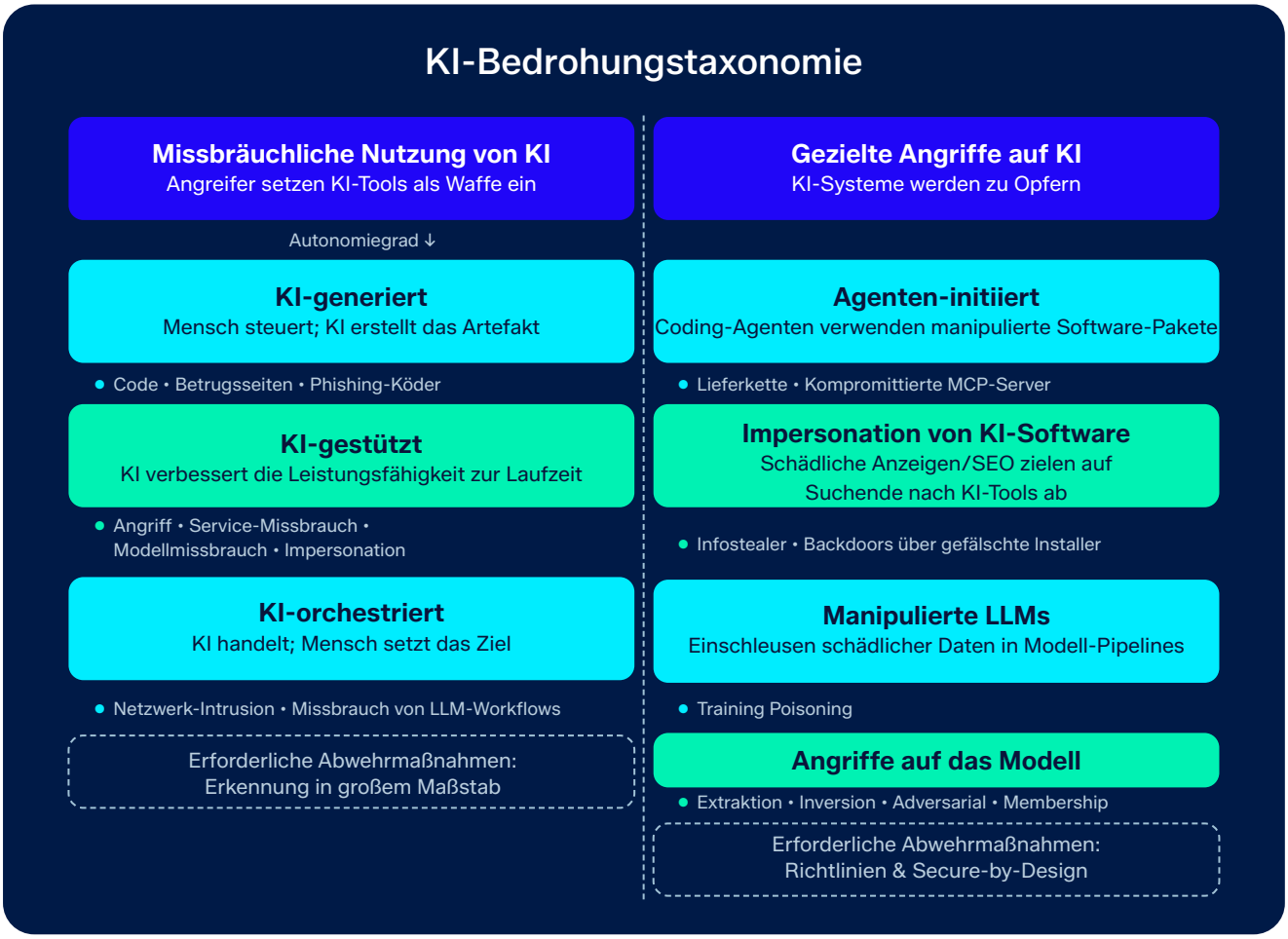


Abbildung 1: Übersicht zur KI-Bedrohungstaxonomie

Autonomiegrad beim KI-Einsatz

Bei der einfachsten Form von KI-generierten Angriffen erstellt ein Mensch mithilfe generativer KI ein Artefakt und setzt dieses manuell ein. Ein Angreifer, der es auf mexikanische Regierungsorganisationen abgesehen hatte, verwendete Claude Code und GPT, um Skripte zu generieren. Durchgesickerte Chats der Ransomware-Gruppe The Gentlemen bestätigten ähnliche Anwendungen. In einem kürzlich vom Sophos IR-Team untersuchten Fall von DragonForce-Ransomware legte der Angreifer während der Verhandlungen mit dem betroffenen Unternehmen einen Report vor, von dem wir stark vermuten, dass er mithilfe von KI erstellt wurde. Dieser enthielt eine detaillierte Analyse der entwendeten Daten, einschließlich personenbezogener Daten und rechtlicher Risiken. Die KI-generierte Analyse wurde im Rahmen eines Versuchs genutzt, Druck auszuüben und das Unternehmen zur Zahlung des Lösegelds zu bewegen.

Etwas mehr Autonomie beim Einsatz von KI beobachten wir bei Angriffen, deren Fähigkeiten zur Laufzeit durch den Einsatz eines KI-Modells verbessert werden. LameHug, eine Python-basierte Malware, die CERT-UA mit mittlerer Zuverlässigkeit der Gruppe IRON TWILIGHT (APT28) zuschreibt, fragte ein gehostetes Open-Weight-Modell ab, um dynamisch umgebungsspezifische Windows-Aufklärungsbefehle zu generieren. Dadurch verlor die statische Analyse erheblich an Aussagekraft und der ausgehende Datenverkehr zu KI-/ML-API-Endpoints wurde zu einem der wenigen verlässlichen Erkennungsmerkmale. KI-gestützte Impersonation ist ein weiterer aktiver Angriffsvektor: Tools zum Klonen von Stimmen und zum Austauschen von Gesichtern werden mittlerweile bei Echtzeitbetrug, Impersonation von Führungskräften und der Umgehung von KYC-Verfahren eingesetzt.

Am äußersten Ende des Spektrums deckte Anthropic im November 2025 eine vom chinesischen Staat finanziell unterstützte Kampagne auf, bei der durch Missbrauch von Claude Code versucht wurde, etwa dreißig Ziele zu kompromittieren. Diese Angriffe sind nach wie vor selten, verkürzen jedoch die Zeit zwischen Aufklärung und Aktion auf eine Weise, die traditionelle Abwehrannahmen infrage stellt.

Case Study: *STAC6994*

Sophos-Analysten beobachteten einen Angreifer, der KI einsetzte, um EDR-Evasionstaktiken in einem „Red Team" Post-Exploitation Framework zu testen, und entdeckten ein schädliches Gerät, das in einer Kundenumgebung kontinuierliche Payload-Generierung durchführte. Der Angreifer hatte virtuelle Maschinen von Ludus bereitgestellt und verwendete die Cursor-IDE mit KI-Agenten, um Post-Exploitation-Tools zu entwickeln und zu testen.

Das Framework führte parallele Tests in drei Umgebungen durch: eine VM mit Sophos Endpoint Protection, eine mit CrowdStrike und eine ohne EDR als Kontrolle. Eine vierte VM fungierte als Sliver-C2-Server. Etwa 12 KI-Agenten agierten mit definierten Rollen.

Ein Agent, der Claude Opus 4.5 ausführte, übernahm die Kernoperationen und die Regelfestlegung. Andere verwalteten die Härtung der operativen Sicherheit (OPSEC), die Dokumentation, Proxy-Stresstests, die Bereitstellung von VMs und das Testen von Tools gegen die EDR-Agenten. Code-Commits flossen über das Model Context Protocol (MCP) an Git.

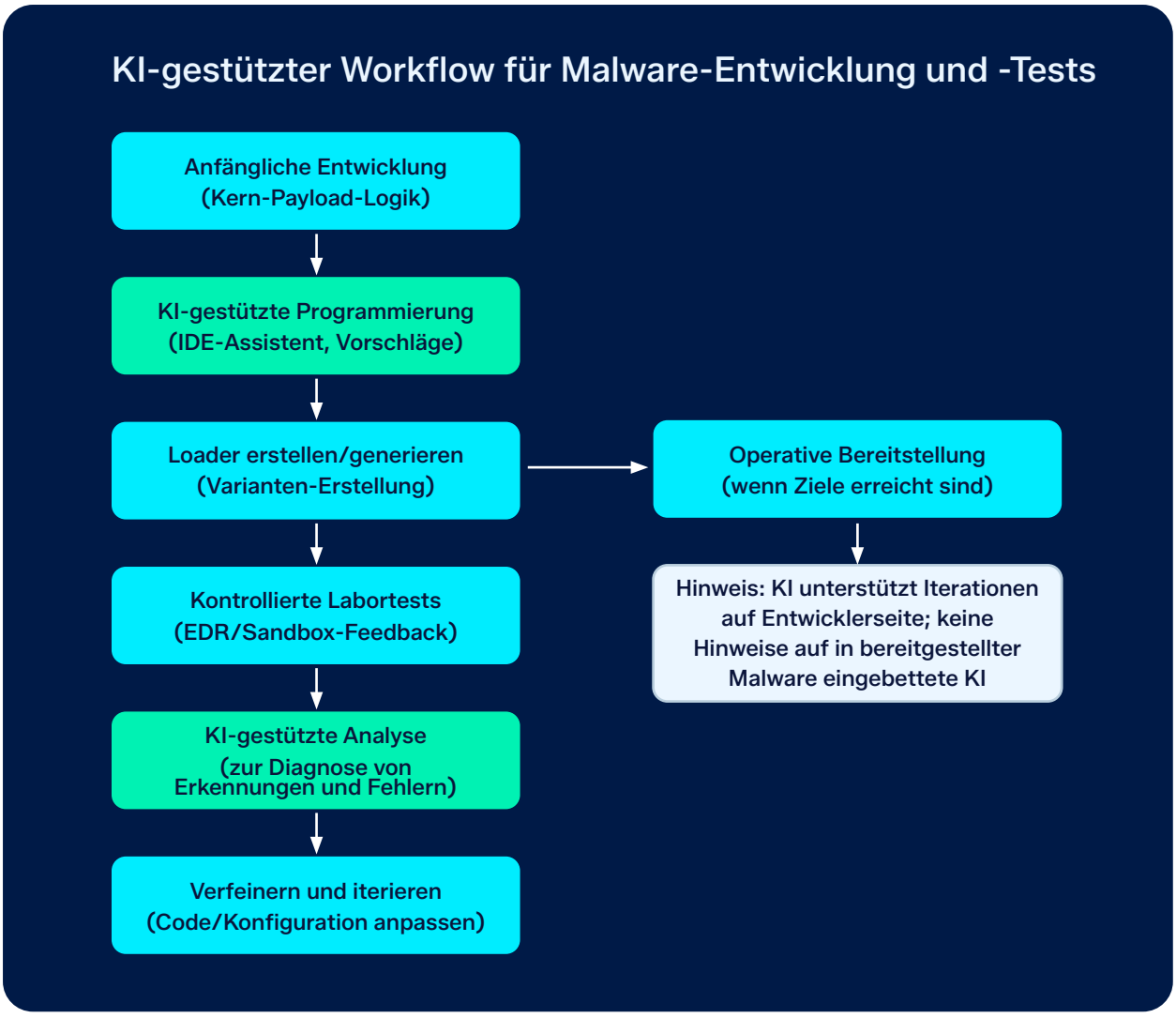


Abbildung 2: Diagramm zur Rolle der KI im Malware-Entwicklungs-Workflow

Das Playbook war systematisch. Den sichergestellten Artefakten zufolge lasen die Agenten Forschungsartikel, die sie aus dem SpecterOps-Blog extrahiert hatten, entnahmen daraus Angriffstechniken, ordneten diese dem MITRE ATT&CK-Framework zu, identifizierten die zur Reproduktion jeder Technik erforderlichen Schritte und Tools, bereiteten die Laborumgebung vor, führten sie aus und berichteten über Ergebnisse.

Im Mittelpunkt stand ein modularer, auf Python basierender Payload-Generator, der unverarbeitete Payloads in Schichten aus Verschlüsselung und Evasionstechniken einbettete und so maßgeschneiderte ausführbare Dateien in Rust und Go erzeugte. Es wurden fast 80 verschiedene Module entwickelt, die über 70 verschiedene Techniken testeten. Der tatsächliche EDR-Umgehungspfad war ein strukturierter technischer Testzyklus mit menschlicher Überprüfung und Iteration. KI koordinierte den Workflow und unterstützte die Experimente; es gab keine Hinweise darauf, dass KI in der bereitgestellten Malware eingebettet war.

Nach mehreren Durchläufen wurde in der Dokumentation der Agenten ein nahezu durchgängiger Erfolg gegenüber den EDR-Agenten attestiert. Sophos-Analysten stellten allerdings fest, dass die Beweislage diese Schlussfolgerung nicht vollständig stützte. Das entscheidende operative Signal war das Tempo: Der gesamte Zyklus komprimierte Wochen auf Tage.

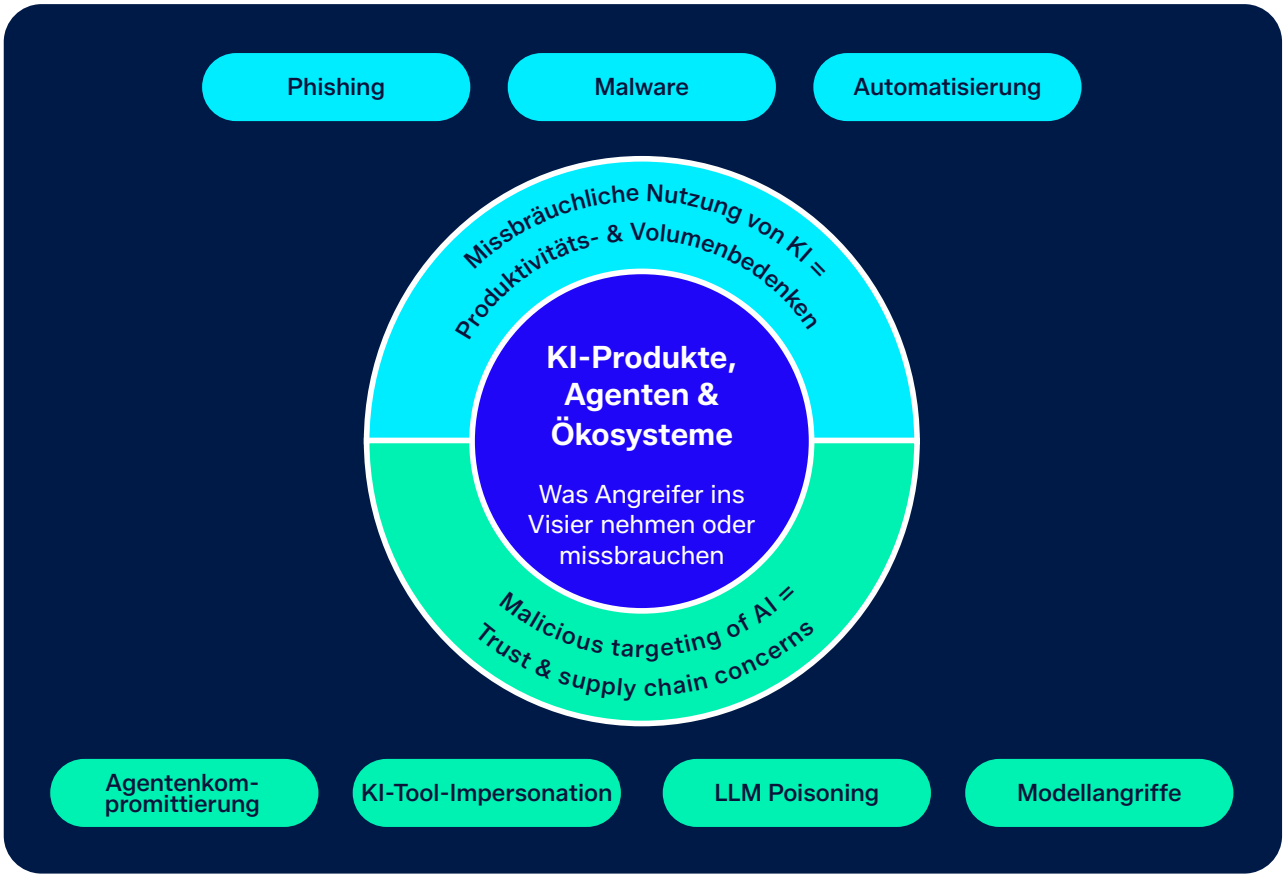
Die CTU bestätigte, dass der STAC6994-Betreiber anschließend an der Bereitstellung von Ransomware und Datendiebstahl beteiligt war; das Framework war ein Produktionstool für echte Angriffe.

Gezielte Angriffe auf KI

Der zweite Teil unserer Taxonomie befasst sich mit Angriffen, bei denen KI-Produkte, -Agenten und -Ökosysteme das Ziel oder die Angriffsfläche sind und nicht das Werkzeug selbst. Darauf werden wir in Teil 2 näher eingehen. Eine durch Agenten initiierte Kompromittierung ist eine der unmittelbar besorgniserregendsten Unterarten: Ein Coding-Agent lädt ein kompromittiertes NPM-Paket herunter oder interagiert bei der Ausführung seiner Aufgabe mit einem manipulierten MCP-Server. Anlass zur Sorge bereitet hier die Verkürzung des Zeitraums zwischen Veröffentlichung und Ausführung, ohne dass ein Mensch in den Prozess eingebunden ist, um das Änderungsprotokoll zu überprüfen.

Unsere Taxonomie umfasst auch die Vortäuschung von KI-Software (bei der Angreifer die Suche von Benutzern nach legitimen KI-Tools ausnutzen), LLM-Manipulation (Konfiguration eines benutzerdefinierten Modells oder absichtliches Einschleusen von schädlichen oder irreführenden Inhalten, um das Verhalten eines Modells zu ändern) sowie verschiedene Angriffe auf die Modelle selbst, z. B. Modellextraktion, Modellinversion, Adversarial Examples und Membership Inference.

Die getrennte Behandlung der beiden Kategorien auf oberster Ebene bietet ein klareres Bild der Bedrohung. Die missbräuchliche Nutzung von KI ist grundsätzlich ein Problem der Produktivität und des Volumens; bestehende Erkennungs-Stacks können zwar vieles erfassen, aber das Problem liegt in der Skalierung und der Iterationsgeschwindigkeit. Gezielte Angriffe auf KI sind ein Problem des Vertrauens und der Lieferkette, das am besten mit Richtlinien, Secure-by-Design Frameworks und ähnlichen Initiativen angegangen werden sollte.



Untergrund-Akzeptanz und -Skepsis

Sophos verfolgt die Einstellungen gegenüber generativer KI in kriminellen Foren seit 2023, u. a auch im Rahmen einer Folgeuntersuchung im Jahr 2025 und im Zuge einer laufenden Beobachtung. Das Bild hat sich erheblich gewandelt. Angreifer kaufen über Zwischenhändler API-Schlüssel für ChatGPT, Claude und Grok, und in Foren sind spezielle KI-Kanäle entstanden, in denen Nutzer Prompt-Vorlagen und „Jailbreaking"-Techniken austauschen. Forscher der Sophos CTU stellten fest, dass ein bekannter Untergrund-Rekrutierer, der zuvor bei der Einstellung von Blockchain-Entwicklern und Call-Center-Mitarbeitenden für Phishing beobachtet wurde, gezielt einen KI-Prompt-Engineer suchte, um KI-Expertise auf die gleiche Weise auszulagern, wie Bedrohungsakteure die Entwicklung von Malware und den Erstzugriff auslagern.

Die CTU beobachtete auch Anzeigen für KI-Sprachbots, die für Vishing und anrufbasierten Betrug entwickelt wurden. Positive Nutzerbewertungen deuten darauf hin, dass die Bots Tonfall, Sprechrhythmus und Gesprächsmuster überzeugend nachahmen können. Der Nutzer „HackingRealm" bewarb einen Service namens „AI OnlyFans Models" für Liebesbetrug und Social Engineering, bei dem synthetische Profilbilder und Gesprächsinhalte generiert wurden, um skalierbare Nutzerprofile auf Dating- und Social-Media-Plattformen zu erstellen – komplett mit einer begleitenden Website und einer Feedback-Seite mit positiven Bewertungen.

Als „KI-gesteuert" vermarktete Tools tauchen auf: Leak Bazaar (ML-gestützte Triage gestohlener Daten) und ein aktualisiertes Cobalt Strike mit REST-API- und MCP-Server-Integration. Angreifer binden LLM-Integrationen in bekannte Workflows ein. Der Verweis auf Claude in der Cobalt-Strike-Werbung soll zum einen Modernität signalisieren, spiegelt aber auch wider, wie sich die LLM-Integration auf kriminellen Marktplätzen zunehmend zu einem Unterscheidungsmerkmal entwickelt – selbst wenn sie eher mehr Komfort als ausgefeilte Vorgehensweisen bietet.

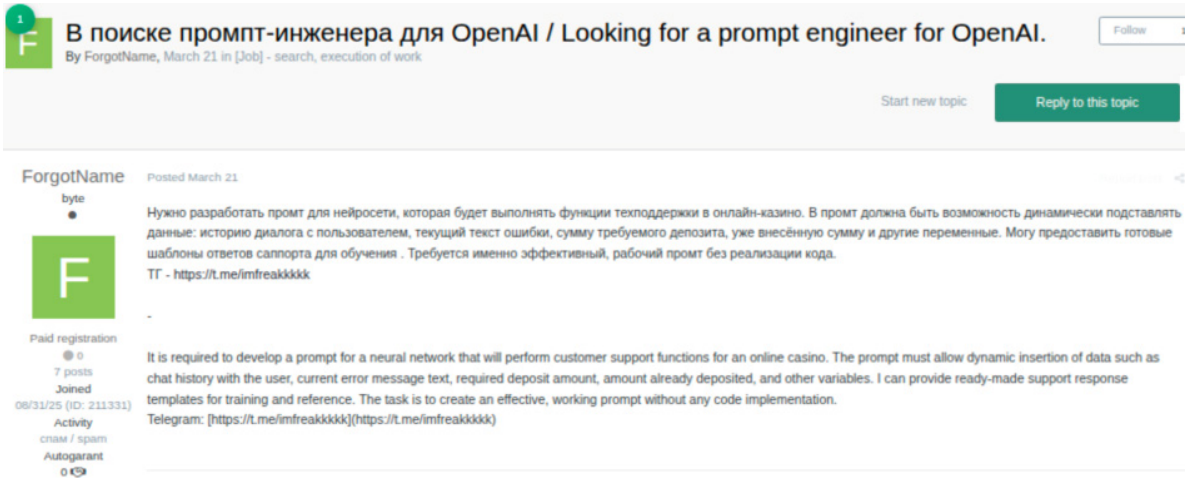
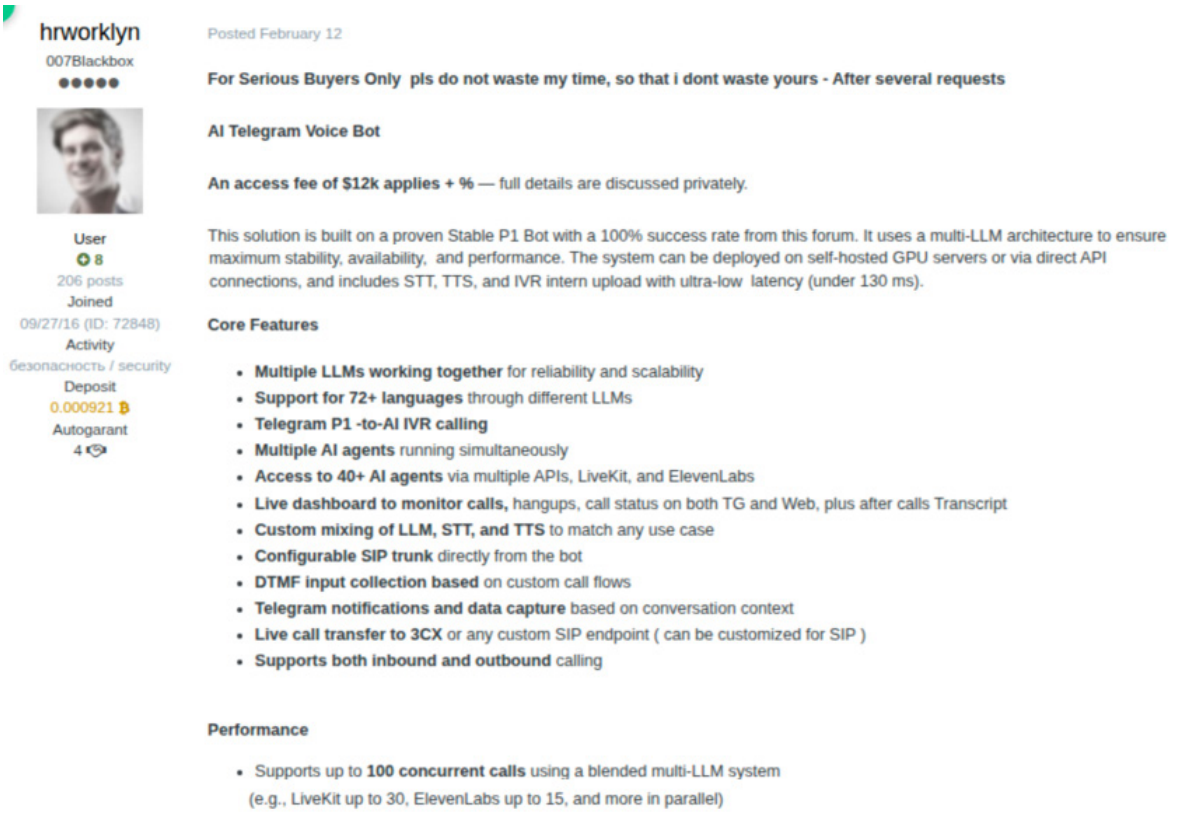


Abbildung 3: Stellenausschreibung für einen OpenAI Prompt Engineer



hrworklyn
007Blackbox
●●●●●

User
8
206 posts
Joined
09/27/16 (ID: 72848)
Activity
безопасность / security
Deposit
0.000921 ₿
Autogarant
4

Posted February 12

For Serious Buyers Only pls do not waste my time, so that i dont waste yours - After several requests

AI Telegram Voice Bot

An access fee of $12k applies + % — full details are discussed privately.

This solution is built on a proven Stable P1 Bot with a 100% success rate from this forum. It uses a multi-LLM architecture to ensure maximum stability, availability, and performance. The system can be deployed on self-hosted GPU servers or via direct API connections, and includes STT, TTS, and IVR intern upload with ultra-low latency (under 130 ms).

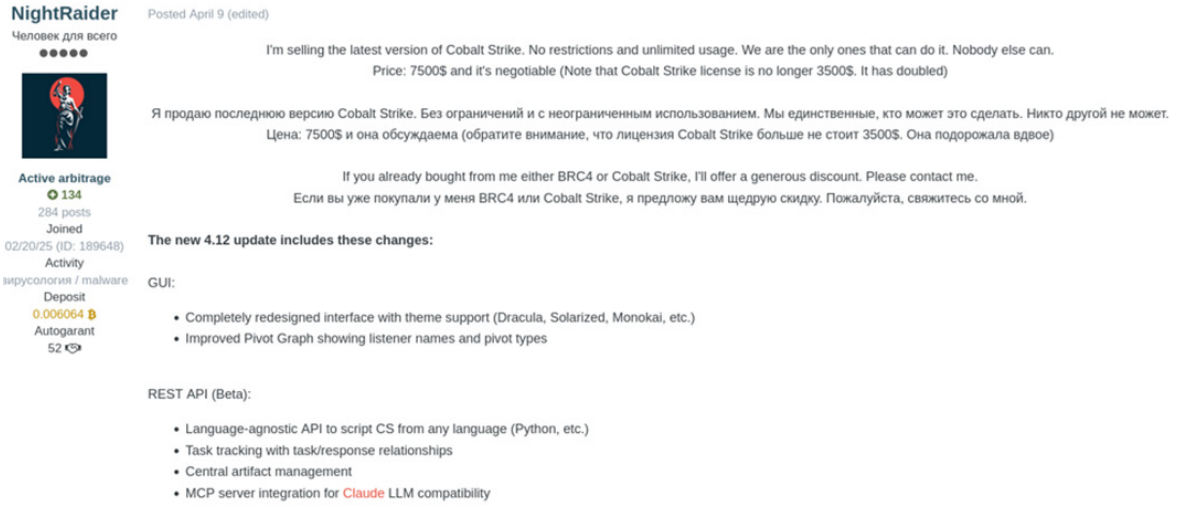
Core Features

- **Multiple LLMs working together** for reliability and scalability
- **Support for 72+ languages** through different LLMs
- **Telegram P1 -to-AI IVR calling**
- **Multiple AI agents** running simultaneously
- **Access to 40+ AI agents** via multiple APIs, LiveKit, and ElevenLabs
- **Live dashboard to monitor calls,** hangups, call status on both TG and Web, plus after calls Transcript
- **Custom mixing of LLM, STT, and TTS** to match any use case
- **Configurable SIP trunk** directly from the bot
- **DTMF input collection based** on custom call flows
- **Telegram notifications and data capture** based on conversation context
- **Live call transfer to 3CX** or any custom SIP endpoint (can be customized for SIP)
- **Supports both inbound and outbound** calling

Performance

- Supports up to **100 concurrent calls** using a blended multi-LLM system (e.g., LiveKit up to 30, ElevenLabs up to 15, and more in parallel)

Abbildung 4: Werbung für einen KI-Telegram-Sprachbot



NightRaider
Человек для всего
●●●●●

Active arbitrage
134
284 posts
Joined
02/20/25 (ID: 189648)
Activity
вирусология / malware
Deposit
0.006064 ₿
Autogarant
52

Posted April 9 (edited)

I'm selling the latest version of Cobalt Strike. No restrictions and unlimited usage. We are the only ones that can do it. Nobody else can.
Price: 7500$ and it's negotiable (Note that Cobalt Strike license is no longer 3500$. It has doubled)

Я продаю последнюю версию Cobalt Strike. Без ограничений и с неограниченным использованием. Мы единственные, кто может это сделать. Никто другой не может.
Цена: 7500$ и она обсуждаема (обратите внимание, что лицензия Cobalt Strike больше не стоит 3500$. Она подорожала вдвое)

If you already bought from me either BRC4 or Cobalt Strike, I'll offer a generous discount. Please contact me.
Если вы уже покупали у меня BRC4 или Cobalt Strike, я предложу вам щедрую скидку. Пожалуйста, свяжитесь со мной.

The new 4.12 update includes these changes:

GUI:

- Completely redesigned interface with theme support (Dracula, Solarized, Monokai, etc.)
- Improved Pivot Graph showing listener names and pivot types

REST API (Beta):

- Language-agnostic API to script CS from any language (Python, etc.)
- Task tracking with task/response relationships
- Central artifact management
- MCP server integration for Claude LLM compatibility

Abbildung 5: Ein Benutzer in einem Cybercrime-Forum bewirbt ein Cobalt-Strike-Update

Doch obwohl es eindeutig Anzeichen für aktive Experimente und die Übernahme dieser Technologie gibt, ist die Community – zumindest in den von der CTU untersuchten Foren – nicht durchweg begeistert. In Anlehnung an frühere Erkenntnisse von Sophos, die im Rahmen einer Untersuchung der Einstellungen gegenüber generativer KI in kriminellen Foren in den Jahren 2023 und 2025 gewonnen wurden, beobachtete die CTU Nutzerprofile, die Bedenken äußerten, dass KI die Beschäftigungsmöglichkeiten verringern könnte, insbesondere im Bereich Malware-Entwicklung und Skripterstellung. Andere lehnten KI rundweg ab und ermutigten dazu, sich auf menschliche Fähigkeiten zu verlassen.

KI-gestütztes Social Engineering und KI-basierte Deepfakes

KI verändert, wie schnell Social-Engineering-Kampagnen skaliert werden können, wie überzeugend sie in verschiedenen Sprachen wirken und wie kostengünstig sie produziert werden können. Im Report „M-Trends 2026" von Mandiant wird beispielsweise ein Wandel von generischem Phishing hin zu LLM-gestütztem, personalisiertem Social Engineering beschrieben, das darauf abzielt, Vertrauen aufzubauen. Dabei stieg Voice Phishing mit einem Anteil von 11 % zum zweithäufigsten anfänglichen Infektionsvektor auf, während E-Mail-Phishing auf nur noch 6 % zurückging. Der MSP Threat Report 2026 von ConnectWise bestätigte den aktiven Einsatz von Deepfake-basierten Betrugsmaschen und LLM-generierten Phishing-Kampagnen und stellte fest, dass KI bereits bestehende Taktiken schneller, skalierbarer und überzeugender machte, anstatt neue Angriffskategorien zu schaffen.

Deepfakes haben sich zu einem operativen Werkzeug für verschiedene Kategorien von Angreifern entwickelt – insbesondere für nordkoreanische Akteure, die mithilfe von KI-generierten Identitäten und Deepfakes Fernbeschäftigungen bei westlichen Unternehmen erlangen und sich so dauerhaften Insider-Zugang verschaffen. Sophos hat ein spezielles CISO-Playbook zu den Taktiken nordkoreanischer IT-Arbeiter veröffentlicht, das Erkennungsindikatoren und organisatorische Gegenmaßnahmen behandelt.

Die Sophos CTU untersuchte ein Sha-Zhu-Pan-Schema, das KI-basiertes Social Engineering beinhaltete. Ein Opfer mit Sitz in Großbritannien wurde über monatelange KI-thematische „Lektionen" in der Messaging-App LINE auf eine gefälschte, KI-gestützte Investment-Plattform namens DAIQ Wealth gelockt. Ein Netzwerk von Betrügern baute über Gruppenchats, die jeweils von einer anderen Person betrieben wurden, eine Vertrauensbasis auf, was auf eine organisierte Gruppe hindeutet, die vordefinierten Skripten folgt. Das Opfer verlor Hunderttausende von Pfund. Als das Opfer Bedenken äußerte, fast 20.000 GBP in bar zu Hause aufzubewahren, organisierten die Betrüger innerhalb von 24 Stunden einen Bargeldkurier zu der Adresse in Großbritannien – komplett mit einer Quittung mit dem Logo eines seriösen Finanzinstituts, die ohne dessen Wissen verwendet wurde.

Zusammenfassung

Wie bereits erwähnt, deuten unsere STAC6994 Case Study, die verdeckte Anwerbung von Prompt Engineers, das Anpreisen von KI als Verkaufsargument bei der Veräußerung von Cybercrime Services sowie Betrugskampagnen zum Thema KI darauf hin, dass Angreifer KI einsetzen, um Angriffe zu ermöglichen und zu ergänzen. Gleichzeitig experimentieren sie mit KI.

Ein ähnliches Bild zeigt sich in der gesamten Branche. Claude unterstützte Angriffe auf Netzwerke der mexikanischen Regierung. Der AI Risk and Resilience Report von Mandiant beschreibt zwei dokumentierte Malware-Familien, die während der Ausführung LLMs abfragen: PROMPTFLUX, ein experimenteller VBScript Dropper, der die Gemini API zur Just-in-Time-Selbstmodifikation verwendet; und PROMPTSTEAL, das IRON TWILIGHT (APT28) zugeschrieben wird und die erste bestätigte Beobachtung von staatlich geförderter Malware darstellt, die in Live-Operationen gegen die Ukraine ein LLM abfragt. Der im Verizon DBIR dokumentierte VoidLink (ein Malware Framework, das von einem KI-Agenten innerhalb von sechs Tagen zusammengestellt wurde) und PromptLock (eine KI-gestützte Ransomware, die erstmals im August 2025 von ESET entdeckt wurde, wobei sich später herausstellte, dass es sich höchstwahrscheinlich um einen akademischen Machbarkeitsnachweis handelte und nicht um ein schädliches, „in-the-wild" zirkulierendes Sample).

Allerdings gibt es bislang keine bestätigten Fälle von vollständig autonomen, KI-gesteuerten Intrusion-Kampagnen in großem Umfang. Im GuidePoint GRIT 2026-Ransomware Report wurde ausdrücklich festgestellt, dass die Bedenken hinsichtlich „Super-Affiliates, die vollständig autonome Ransomware-Bots einsetzen", übertrieben seien und dass der Einsatz von KI und LLM bei weniger erfahrenen Angreifern nach wie vor nur in rudimentärer Form stattfinde. Das AIM3 Assessment von Recorded Future stufte den Großteil der beobachteten KI-Malware in ihrem KI-Malware-Reifegradmodell auf die Reifegrade 1–3 ein (Experimentieren, Einführen, Optimieren). Dabei wurden keine bestätigten Beispiele für wirklich eingebettete „Bring-your-own-AI"-Malware gefunden, die lokale Modelle auf den Hosts der Opfer ausführt, und es wurde der Schluss gezogen, dass „Verteidiger den Missbrauch legitimer KI-Services-überwachen, bestehende Kontrollen verstärken und Bedrohungen den AIM3-Stufen zuordnen sollten, anstatt auf Science-Fiction-Szenarien überzureagieren."

Der von Anthropic dokumentierte Fall GTG-1002 (Claude Code, der 80–90 % einer Datendiebstahlkampagne mit mehreren Zielen gegen etwa 30 Organisationen automatisierte) kommt einem autonomen Betrieb im öffentlichen Raum am nächsten, stellt jedoch eher eine Ausnahme dar und ist kein Beispiel für einen Trend.

Teil 2:
KI-Risiko in Unternehmen

KI hat bereits Einzug in Unternehmen gehalten: Programmieragenten schreiben und implementieren Code mit den Zugangsdaten von Entwicklern, agentenbasierte Assistenten lesen E-Mails und rufen APIs auf, zahlreiche Systeme werden miteinander vernetzt und interne Teams experimentieren mit Open-Weight-Modellen in Managed-Cloud-Infrastrukturen. Open-Source- und Open-Weight-Modelle werden immer leistungsfähiger, kosten in der Regel nur einen Bruchteil von Closed-Source- und Closed-Weight-Modellen und ermöglichen es Unternehmen, sich Outputs direkt zu eigen zu machen und die Kontrolle über ihre Daten zu behalten. Allerdings stammen diese Modelle größtenteils aus der Volksrepublik China. Die Governance-Frage lautet nicht mehr, ob die Einführung erlaubt werden soll, sondern wie sie sicher gestaltet werden kann, bevor sich die Angriffsfläche aufgrund schlechter Standardeinstellungen verfestigt.

Selbst risikoscheue Unternehmen, die KI in genehmigten Anwendungsfällen nur in geringem Umfang einsetzen, haben aller Wahrscheinlichkeit nach ein Problem mit „Schatten-KI" – und wie wir gleich noch erläutern werden, sind die Befürchtungen hinsichtlich Schatten-KI und Datenlecks durchaus berechtigt. Wir werden die Einzelheiten zu Risikotoleranz, Kontrollen und Transparenz in Teil 4 besprechen.

KI-Entwicklungsinfrastruktur als Angriffsziel

Ein deutlicher Trend im Jahr 2026 ist die gezielte Ausrichtung auf KI-spezifische Entwicklungsinfrastruktur und insbesondere auf die Vertrauensbeziehungen, die Entwickler um KI-Tools herum aufbauen. Im Gegensatz zu den meisten KI-bezogenen Bedrohungen, die noch im Entstehen begriffen oder weitgehend auf theoretische Machbarkeitsnachweise beschränkt sind, ist dies der Bereich, in dem Angriffe derzeit stattfinden.

Ein deutlicher Trend im Jahr 2026 ist die gezielte Ausrichtung auf KI-spezifische Entwicklungsinfrastruktur und insbesondere auf die Vertrauensbeziehungen, die Entwickler um KI-Tools herum aufbauen.

Die **SANDWORM_MODE** npm-Wurm-Kampagne umfasste mindestens 19 Typosquatting-Pakete, die darauf ausgelegt waren, legitime Entwickler-Dienstprogramme und KI-Codierungstools nachzuahmen. Die Pakete installierten einen manipulierten MCP-Server und zwangen legitime KI-Assistenten durch eingebettete Prompt Injection dazu, SSH-Schlüssel und Cloud-Zugangsdaten unbemerkt abzurufen und ohne Wissen des Benutzers zu exfiltrieren.

Bei dem Angriff auf die **VS-Code-Erweiterung „Nx Console"** wurden Zugangsdaten von HashiCorp Vault, npm, AWS, GitHub, 1Password sowie API-Schlüssel von Anthropic gestohlen. Der Angriff schien Teil der umfassenderen TeamPCP-/mini-Shai-Hulud-Kampagne zu sein. Eine kompromittierte Erweiterung innerhalb der IDE eines Entwicklers hat direkten Zugriff auf die Zugangsdaten und Repositorys, die am wichtigsten sind.

Das **Anthropic Claude Code-Quellcode-Leck** im März 2026, bei dem Quellcode versehentlich in einem npm-Paket veröffentlicht wurde, veranschaulicht eine andere Dimension. Der Code enthielt Berichten zufolge etwa 1.900 Dateien und 500.000 Codezeilen, einschließlich Details zu unveröffentlichten Funktionen. Zwar gab Anthropic an, dass keine Kundendaten offengelegt wurden, doch könnte eine Analyse des Codes die künftige Identifizierung von Fehlern und Sicherheitslücken ermöglichen. KI-Tools selbst sind zu einem Intelligence-Ziel geworden.

Unsere Empfehlung: Jedes Unternehmen und jede Organisation mit einem Entwicklerteam sollte dies als ihr größtes und dringlichstes Risiko betrachten. Die Kontrollen hier sind weitgehend prozessgesteuert – sorgfältiges Abhängigkeitsmanagement, Version-Pinning und die Sicherstellung von Überprüfungen neuer Open-Source Libraries vor der Einbindung – gepaart mit einer genauen Überwachung der Endpoints von Entwicklern. (In Teil 4 stellen wir eine Liste mit sieben Maßnahmen vor, die Verteidiger bereits jetzt ergreifen können, um den Schadensradius bei der Bereitstellung agentischer KI zu verringern.)

Wo sich die Gefährdung konzentriert

Die Identitätsstruktur, die KI-Services mit Unternehmenssystemen verbindet, schafft eine Gefährdung, für deren Bewältigung die bestehende Governance nicht ausgelegt ist. IBM X-Force behauptete, dass im Jahr 2025 über 300.000 ChatGPT-Zugangsdaten im Darknet zum Kauf angeboten wurden und dass in einem Forumsbeitrag vom Februar 2025 festgestellt wurde, dass über 20 Mio. ChatGPT-Konten gestohlen worden seien (diese Zahl bleibt jedoch bislang unbestätigt). Der Salesloft/Drift-Vorfall demonstrierte einen konkreten Mechanismus: Mithilfe von Drift-OAuth-Token wurden mehrere Salesforce-Umgebungen kompromittiert, was verdeutlicht, wie der tokenbasierte Zugriff eines Chatbots unmittelbar zu einer Systemkompromittierung führen kann.

BeyondTrust meldete einen Anstieg von 466,7 % bei aktiven KI-Agenten in Unternehmensumgebungen im vergangenen Jahr. Diese Agenten führen bestimmte Angriffsflächen ein: Prompt Injection, Aufruf schädlicher Tools und manipulierte Dateneingaben. Microsofts eigene Copilot-Schwachstelle CVE-2025-32711 (EchoLeak) demonstrierte, wie eine unzureichende Eingabevalidierung Zero-Click-Datenlecks ermöglichen kann. Wenn Maschinenidentitäten keine MFA-Äquivalente besitzen, langfristig gültige Geheimnisse mit sich führen und mit übermäßigen Berechtigungen agieren, werden sie zu einem realistischen und attraktiven Angriffsziel.

Angreifer sind sich dieser Möglichkeiten sehr wohl bewusst und erkunden aktiv die Angriffsfläche. Im Januar 2026 beschrieb GreyNoise Kampagnen, die LLM-Bereitstellungen kartierten. Forscher dokumentierten Operation Bizarre Bazaar: eine Kampagne, bei der ungeschützte LLM- und MCP-Endpoints gekapert, validiert und der Zugriff darauf über eine mit silver.inc verbundene Infrastruktur weiterverkauft wurde.

Die Governance-Lücke

In der gesamten Branche und insbesondere in den Führungsetagen herrschen sowohl Bedenken hinsichtlich der sicherheitsrelevanten Auswirkungen der KI als auch ein Mangel an Kapazitäten und Kompetenzen, um diesen Auswirkungen zu begegnen.

Die Identitätsstruktur, die KI-Services mit Unternehmenssystemen verbindet, schafft eine Gefährdung, für deren Bewältigung die bestehende Governance nicht ausgelegt ist.

ChatGPT allein generierte 410 Mio. DLP-Richtlinienverstöße in Zscalers Daten. Der Saviynt/Cybersecurity Insiders CISO AI Risk Report 2026 ergab, dass 75 % der Befragten Schatten-KI-Tools entdeckt hatten und 95 % bezweifelten, dass sie Missbrauch erkennen oder eindämmen könnten. Nur 1 % der Unternehmen verfügt über ein dediziertes KI-Sicherheitsbudget (Pentera), und laut Splunks CISO Report 2026 herrscht zwar Konsens darüber, dass KI bei der Produktivität und der Bewältigung großer Datenmengen helfen kann, doch eine Mehrheit der CISOs ist besorgt über Datenlecks, Schatten-KI und Halluzinationen. Interessanterweise stellte Splunk fest, dass diese Ängste bei CISOs in Unternehmen mit höherer KI-Reife ausgeprägter waren, und deutete an, dass diejenigen, „die auf ihrer generativen KI-Reise bereits weiter fortgeschritten sind, allmählich gewisse Kompromisse bemerken."

Die Einführung von KI schreitet schneller voran als die dafür vorgesehenen organisatorischen Prozesse – und die regulatorischen Rahmenbedingungen werden immer strenger.

Diese Kompromisse und Ängste, insbesondere im Hinblick auf Schatten-KI, sind durchaus begründet.
Im April 2026 legte Vercel eine Sicherheitsverletzung offen, die durch die Nutzung des KI-Produktivitätstools Context.ai durch einen Mitarbeiter ausgelöst wurde. Der Mitarbeiter registrierte sich mit seinem Google Workspace-Unternehmenskonto und erteilte der OAuth-App von Context.ai die Berechtigung „Alle zulassen". Als Context.ai kompromittiert wurde – Berichten zufolge durch eine Lumma Stealer-Infektion auf dem Gerät eines Context.ai-Mitarbeiters –, übernahmen die Angreifer diese OAuth-Token und drangen in die internen Umgebungen von Vercel ein.

Die Governance-Herausforderung besteht darin, dass die Einführung von KI schneller voranschreitet als die dafür vorgesehenen organisatorischen Prozesse – und dass sich das regulatorische Umfeld verschärft, während die Kluft zwischen der Einführungsgeschwindigkeit und Governance-Reife wächst.
Die europäische KI-Verordnung schreibt vor, dass risikoreiche Systeme ab 2026 die Bestimmungen vollständig einhalten müssen; bei Verstößen drohen Strafen in Höhe von bis zu 7 % des weltweiten Umsatzes. Die Transparenzanforderungen des US-Bundesstaats Kalifornien traten am 1. Januar 2026 in Kraft. Saviynt/Cybersecurity Insiders stellte jedoch fest, dass zwar 71 % der Großunternehmen KI-Agenten implementiert haben, die auf geschäftskritische Systeme zugreifen, aber nur 16 % diesen Zugriff wirksam regeln.

Als erster Schritt zum Schließen dieser Lücke sollte der Prompt-Verkehr als Sicherheitsprotokoll höchster Priorität behandelt werden: Durch die Erfassung und Analyse von Prompts lassen sich Injection-Versuche, Datenexfiltration, Missbrauch durch Insider sowie unerwartetes Agentenverhalten aufdecken. Dies kann bei cloudbasierten Modellen auf der Proxy-Ebene oder bei lokal bereitgestellten Agenten auf der Harness-Ebene beginnen, ohne dass auf die vollständige Modellintegration gewartet werden muss. Doch Prompts sind nur ein Teil der Antwort; Tools und Fähigkeiten werden zu einem grundlegenden Teil der Angriffsfläche und sollten auch als solche behandelt werden.

Teil 3:
Wie KI defensive Betriebsmodelle verändert

Triage und Analyse

Bei der Umsetzung von Playbooks, der Rekonstruktion von Zeitabläufen und beim Reporting in natürlicher Sprache können Reasoning-Modelle einen echten operativen Mehrwert schaffen. So verarbeitete beispielsweise die MDR-Infrastruktur von Kaspersky im Jahr 2025 rund 400.000 Alerts, wobei eine KI-gestützte Erkennungslogik False Positives bereits vor der Überprüfung durch Analysten herausfilterte; 39.000 Alerts wurden zur weiteren Analyse eskaliert. KI-Modelle sind zunehmend in der Lage, diese Analysen durchzuführen; Prophet Security hob im März 2026 zwei Fälle in The Hacker News hervor, bei denen Modelle kontextbezogene Analysen und Analysen von Vorfällen durchführten, bei denen schädliche Aktivitäten nicht sofort offensichtlich waren. Im ersten Fall ging es um einen inaktiven Benutzer mit einem alten API-Schlüssel, der plötzlich wieder aktiv wurde; im zweiten Fall wurde eine schädliche Absicht in einer Phishing-E-Mail aufgedeckt, die die meisten, wenn nicht sogar alle Sicherheits- und Validierungsprüfungen bestanden hätte. In beiden Fällen, so merkt Prophet Security an, war kein einzelner Datenpunkt für sich genommen verdächtig; die Bedrohung wurde erst durch eine kontextbezogene Multi-Point-Analyse deutlich – die Art von Analyse, die menschliche Ermittler durchführen, wenn sie die Kapazitäten dafür haben. Das Problem ist natürlich, dass Analysten diese Kapazitäten oft nicht haben, und genau hier können Reasoning-Modelle einen Beitrag leisten, indem sie jedem Alert das gleiche Maß an Aufmerksamkeit bei der Analyse zukommen lassen.

KI ist jedoch kein Allheilmittel und erfordert eine sorgfältige technische Umsetzung. Untersuchungen zum Thema „Überdenken" zeigen, dass ab einem bestimmten Schwellenwert zusätzliches Nachdenken dazu führt, dass Modelle richtige Antworten in falsche umkehren. Negative Antwortumkehrungen übertrafen die positiven nach etwa 7.000 Token und erreichten bei 12.000 ein Verhältnis von drei zu eins. SOC-Implementierungen sollten daher den Aufwand für Schlussfolgerungen, die Tiefe von Wiederholungsversuchen und Tool-Call Loops als anpassbare technische Parameter behandeln.

Im Großen und Ganzen verläuft die Einführung uneinheitlich, und die Erwartungen müssen angepasst werden. Dem Splunk CISO Report zufolge setzen derzeit nur 40 % der CISOs generative KI in ihren Sicherheitsfunktionen ein. Der Anteil aktiver Bereitstellungen von agentischer KI liegt derzeit nur bei 6 %; allerdings spielen 39 % mit dem Gedanken, einer Einführung.

Hinzu kommt, dass KI-Tools, die für Verteidiger gedacht sind, ihre eigene Angriffsfläche mit sich bringen können. Der Cobalt State of Pentesting Report kam beispielsweise zu dem Ergebnis, dass KI- und LLM-Anwendungen 2,7-mal so häufig risikoreiche Schwachstellen aufweisen wie herkömmliche Software, wobei nur 38 % der Probleme während der Penetrationstests behoben wurden.

Das Speicherproblem

Die meisten Produktiveinsätze stecken in einer Phase fest, die manche Forscher als zweite oder dritte Generation der Agentenreife bezeichnen: toolgestützte Assistenten, die bei jeder Sitzung von vorne beginnen. Sie behalten ihr Training bei, können sich aber nicht daran erinnern, dass dasselbe Alert-Muster gestern auftrat und sich als harmlos herausstellte oder dass das Blockieren eines bestimmten IP-Bereichs im letzten Monat das VPN unterbrach.

Der Unterschied zwischen einem Senior-Analysten und einem Junior-Analysten liegt nicht unbedingt in der Fähigkeit zum logischen Denken, sondern darin, was der jeweilige Analyst jeweils in eine Situation einbringt: episodische Erinnerungen an vergangene Ereignisse, semantisches Wissen über die Umgebung und prozedurale Intuitionen darüber, was funktioniert. Derzeitige KI-Assistenten agieren eher auf Junior- als auf Senior-Level und analysieren jeden Alert so, als wäre es das erste Mal. Forscher untersuchen Memory-as-Operating-System-Architekturen, External Memory-as-Middleware-Services und Knowledge-Graph-Ansätze mit bitemporaler Schlussfolgerung. Bis diese ausgereift sind, bleiben KI-Systeme nützliche Assistenten und keine autonomen Akteure.

Die Erkennungsgrenze

Verhaltenserkennungssysteme arbeiten mit Live-Telemetrieströmen bei extremem Klassenungleichgewicht. Eine Plattform, die täglich Billionen von Ereignissen verarbeitet, fasst diese zu Millionen deduplizierter Erkennungen zusammen, priorisiert sie zu Tausenden von Alerts mit hohem Schweregrad und generiert täglich Hunderte von Analysen, die auch nach der Überprüfung durch Menschen weiterhin kritisch bleiben. Das End-to-End-Verhältnis liegt bei etwa eins zu zehn Milliarden. Stefan Axelssons wegweisende Arbeit zu Prävalenzfehlern auf dem Gebiet der Intrusion-Erkennung formalisierte die mathematischen Zusammenhänge: Selbst ein Detektor mit einer Genauigkeit von 99,9 % löst bei jeder echten Bedrohung Millionen von Fehlalarmen aus, wenn die A-priori-Wahrscheinlichkeit eines Angriffs ausreichend gering ist.

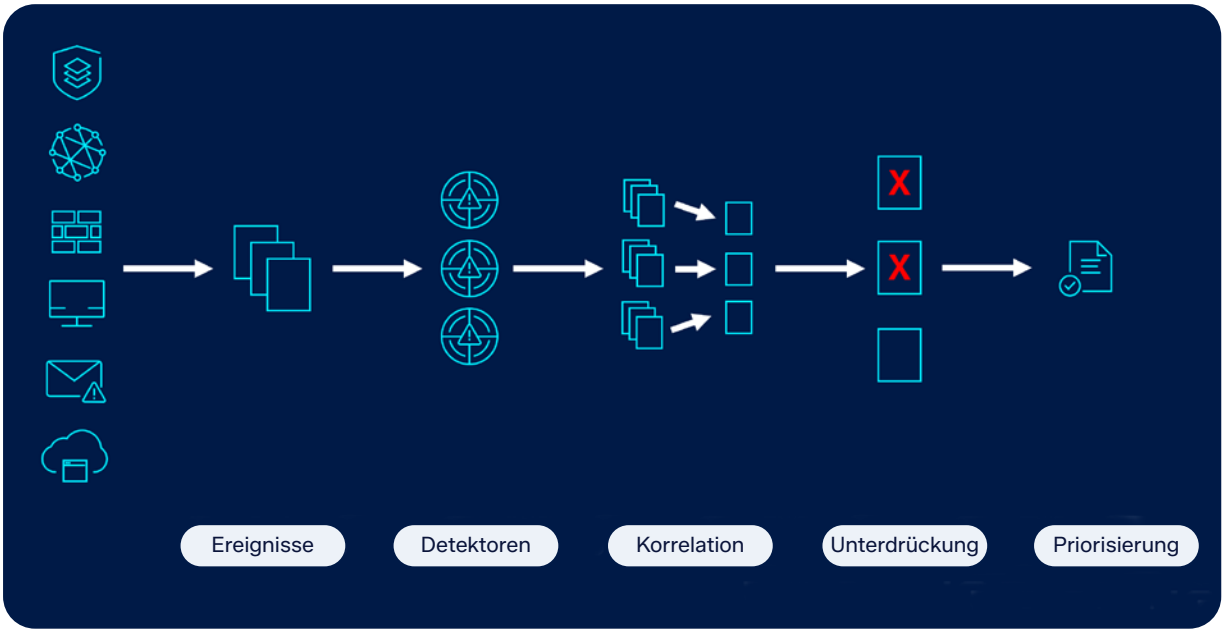


Abbildung 6: Übersicht über die mehrstufige Pipeline

Sophos-Forschungsergebnisse, die auf der NorthSec 2026 vorgestellt wurden, verdeutlichen, warum mehrschichtige Erkennung nach wie vor wichtig ist. Ein zweiwöchiges Telemetriefenster mit 11,8 Billionen Ereignissen wurde mittels zunehmend komplexer Detektoren (einfacher Indikatorabgleich, Korrelationsregeln, zweckgebundene ML-Modelle, Deduplizierung, Korrelation und Unterdrückung) auf 81.573 hohe und kritische Alerts reduziert – durchschnittlich weniger als 50 pro Organisation. Innerhalb dieser verbleibenden Alerts identifizierte Sophos einen Infostealer-Angriff, der mit der TamperedChef-Kampagne in Verbindung stand.

Reasoning-Modelle sind über dieser Ebene angesiedelt. Sie generieren Analysehypothesen, rufen Telemetriedaten ab, fassen sie zu Beweiszusammenfassungen zusammen und koordinieren begrenzte Reaktionsmaßnahmen unter Einhaltung klar definierter Rahmenbedingungen. Sie ersetzen keine Modelle, die mit proprietären Vorfalldaten, umgebungsspezifischen Baselines und kontinuierlichen Feedbackschleifen trainiert wurden, auf die allgemeine Reasoning-Systeme keinen Zugriff haben.

Die McKinsey AI Trust Maturity Survey hat die organisatorischen Hindernisse quantifiziert: Für fast zwei Drittel der Befragten sind Sicherheitsbedenken das größte Hindernis für den großflächigen Einsatz agentischer KI. Ungenauigkeit (74 %) und Cybersicherheit (72 %) sind die führenden spezifischen Risiken. Nur etwa 30 % der Unternehmen/Organisationen haben den Reifegrad 3 („Es werden Schritte unternommen, um alle notwendigen verantwortungsvollen KI-Praktiken zu entwickeln“) oder höher in der KI-Governance erreicht. Unternehmen und Organisationen, in denen die Zuständigkeit für verantwortungsvolle KI klar festgelegt ist, erreichen einen durchschnittlichen Reifegrad von 2,6, während dieser bei Unternehmen und Organisationen ohne klare Zuständigkeiten bei 1,8 liegt.

Die größte Herausforderung für die KI-gestützte Abwehr ist die organisatorische Bereitschaft. Die Modelle sind leistungsfähig genug; die Frage ist, ob Unternehmen und Organisationen sie auf dem Vertrauensniveau betreiben können, das autonome Security Operations erfordern.

Sophos KI-Forschung: Verteidigung des KI-Layers

Neben den oben beschriebenen Forschungsarbeiten zur mehrschichtigen Erkennung hat das Sophos AI Team mehrere weitere neuartige Abwehrtechniken entwickelt, die jeweils auf einen bestimmten Fehlermodus abzielen.

- **Der Untrusted Data Scanner (UDS)** inspiziert externe Inhalte auf Prompt Injection, bevor sie ein LLM oder einen Agenten erreichen. Jeder LLM-Anwendungsfall verfügt über einen Befehlskanal (was die Anwendung dem Modell mitteilt) und einen Datenkanal (den nicht vertrauenswürdigen Inhalt, den es verarbeitet). Prompt Injection schmuggelt Anweisungen in den Datenkanal. UDS ist bewusst auf dieses spezifische Problem ausgerichtet und nicht als allgemeiner Jailbreak-Detektor konzipiert. Erste Ergebnisse zeigen eine niedrige False-Positive-Rate, was von Bedeutung ist, da Sicherheitsdaten voller Fachbegriffe aus dem Sicherheitsbereich sind – Incident Reports, Malware-Analysen, Ergebnisse von Penetrationstests – und ein Detektor, der bei jedem Bedrohungsbericht „falschen Alarm“ auslöst, unbrauchbar wäre.
- **LLM Scoping** klassifiziert, ob Anfragen innerhalb der beabsichtigten Funktion liegen, und weist Anfragen außerhalb des Geltungsbereichs zurück, bevor sie das LLM erreichen. Bei allen untersuchten Methoden schnitten Scoping-Modelle deutlich besser ab als auf System-Prompts basierende Abwehrmechanismen und senkten die Erfolgsraten von Angriffen in mehreren öffentlichen Benchmarks auf nahezu null. Das Team wandte zudem Multi-round Automatic Red Teaming (MART) an, um Scoping-Modelle gegen Anpassungen von Angreifern zu härten.

- **CerBERTus**, ein „dreiköpfiges“ BERT-basiertes Modell, adressiert die Anfälligkeit der binären Jailbreak-Erkennung. Ein einzelner gemeinsamer Encoder speist drei „Klassifizierungsköpfe“: Schädlichkeit (primär), Zielkategorie (was der Benutzer zu tun versucht) und Framing-Stil (wie die Anfrage präsentiert wird). Die Hilfsaufgaben fungieren als induktiver Bias und fördern eine Repräsentation, die das eigentliche Ziel vom Wrapper trennt. Das Team erstellte ein strukturiertes, faktorielles Prompt-Korpus, in dem Ziele (Cyberangriffe, Betrug, Sprengstoffe, Drogensynthese, Datenschutz, extremistische Propaganda und mehr) mit Adversarial Frames (Rollenspiel, Fiktion, Dringlichkeit, akademischer Vorwand, Verschleierung) verknüpft wurden, um diese Trennung zu trainieren und einem Stresstest zu unterziehen.
- **Erkennung von Anomalien**: In einer auf der Black Hat USA 2025 vorgestellten Forschungsarbeit kombinierte das Team die Anomalieerkennung mit LLMs und stellte fest, dass der Erfolg der Methode darauf beruhte, vielfältige harmlose Befehlszeilen zu generieren, und nicht darauf, schädlich Befehlszeilen aufzuspüren – was die False-Positive-Rate bei Befehlszeilen-Klassifikatoren signifikant reduzierte.
- **LLM Salting**, vorgestellt auf der CAMLIS 2025, lässt sich vom Passwort-Salting inspirieren: Es führt gezielte Verhaltensvariationen ein, sodass vorberechnete Jailbreaks nicht über homogene LLM-Bereitstellungen hinweg wiederverwendet werden können.

Cyber-Deception gegen autonome Angreifer

Zwei Jahrzehnte lang wurde Cyber-Deception nie zu einer primären Kontrollmaßnahme. Gegen autonome Angreifer ändert sich das Kalkül. KI-Agenten werden nicht müde und können jeden Pfad mit maschineller Geschwindigkeit durchgehen. Täuschende Umgebungen, wie sie beispielsweise im Palisade LLM Agent Honeypot und im MANTIS-Framework vorgeschlagen werden, können den Kontext eines Agenten mit irreführenden Informationen füllen, sein Inferenzbudget verschwenden und echte wirtschaftliche Kosten verursachen. Ein solches Beispiel, HoneyTrap, zeigte einen Anstieg des „Ressourcenverbrauchs der Angreifer“ um 149 % im Vergleich zu herkömmlichen Abwehrmaßnahmen. Das Gebiet steckt noch in den Kinderschuhen, aber genau die Bedingungen, unter denen Angreifertäuschungen in den Vordergrund rücken, sind jene, die durch die zunehmende Verbreitung offensiver KI geschaffen werden.

Wo die Tempo-Asymmetrie zuschlägt

Ein Angreifer, der KI einsetzt, muss sich weder mit Beschaffungsprozessen noch mit Compliance-Prüfungen oder einem Change Advisory Board auseinandersetzen. Stattdessen kann er ein Tool direkt implementieren und schon innerhalb von wenigen Tagen die von ihm gewünschten Ergebnisse erzielen. Die KI-gestützte Kampagne gegen FortiGate-Geräte kompromittierte in fünf Wochen über 600 Firewalls in 55 Ländern – das Werk eines Angreifers, der kommerzielle KI-Tools einsetzte.

Aufseiten der Verteidigung wird eine Erkennung ausgelöst und eine KI-gestützte Analyse liefert innerhalb von Minuten eine Empfehlung zur Eindämmung, doch der Behebungsprozess des Kunden erfordert ein Änderungsticket, ein Wartungsfenster sowie die Einhaltung einer 48-stündigen SLA-Frist. Die Analyse ist schneller geworden, die Reaktion jedoch nicht. Dies war möglicherweise bei zwei Schwachstellen der Fall, die im vergangenen Jahr schnell ausgenutzt wurden. Wie bereits erwähnt, wurde CVE-2026-10520 (Ivanti Sentry) innerhalb von 24 Stunden nach der Proof-of-Concept-Veröffentlichung ausgenutzt und CVE-2026-42208 (LiteLLM) innerhalb von 36 Stunden.

Während KI-gestützte Triage Verteidigern zweifellos beim Tempo hilft, kann im Voraus geleistete Arbeit an Sicherheitsgrundlagen dazu beitragen, die Asymmetrie hier anzugehen: Verringerung der Angriffsfläche, Durchsetzung von MFA und Aufrechterhaltung einer Telemetrie, die vollständig genug ist, um zu rekonstruieren, was passiert ist, wenn die Prävention fehlschlägt.

Teil 4:
Was Security-Verantwortliche tun sollten

Aus den in diesem Report dargelegten Erkenntnissen lässt sich eine geringe Anzahl bedingter Entscheidungen ableiten, die danach gegliedert sind, wo ein Unternehmen/ eine Organisation derzeit steht.

Kalibrierung der Risikotoleranz für die KI-Einführung

Die in diesem Report dargelegten Fakten erzeugen ein Spannungsfeld, für das Unternehmen und Organisationen eine Lösung finden müssen: Die Einführung von KI birgt echte Risiken, aber eine langsame Einführung ist ebenfalls risikobehaftet. Wenn man zu vorsichtig vorgeht, während die Angreifer an Tempo zulegen, gerät man bei der Abwehr ins Hintertreffen – und genau das ist laut diesem Report die entscheidende Herausforderung. Risiken einzugehen, ist also unvermeidlich; die Frage ist, wie man die richtigen Risiken eingeht.

Das interne Risiko-Framework von Sophos unterscheidet zwischen guten und schlechten Risiken anhand zweier Achsen: Aufwärtspotenzial und Begrenzung von Abwärtsrisiken. Ein gutes Risiko hat spürbare Auswirkungen, Maßnahmen zur Begrenzung des Schadensradius, einen umkehrbaren Pfad (Pilotprojekt, Rollback oder Fallback), klare Verantwortlichkeiten und definierte Feedback-Schleifen.
Ein „schlechtes" Risiko birgt ein unbegrenztes Schadenspotenzial, verstößt gegen unverhandelbare Vorgaben (Sicherheit, Datenschutz, rechtliche Vorgaben, Compliance), verfügt über keinen Wiederherstellungsplan, weist unklare Zuständigkeiten auf oder umgeht die menschliche Beurteilung bei Änderungen, die sich direkt auf Kunden auswirken.

Speziell im Hinblick auf KI mündet dieses Framework vor jeder Bereitstellung in praktischen Fragen: Was ist das schlimmste denkbare Ergebnis? Wie groß ist der potenzielle Schadensradius? Kann der Schritt rückgängig gemacht werden? Wer trägt die ultimative Verantwortung? Bereitstellungen mit begrenztem Risiko, wie beispielsweise ein schreibgeschützter Pilotversuch eines agentenbasierten Triage-Tools in einem isolierten Datensatz, sollten konsequent vorangetrieben werden. Bereitstellungen mit hohem Risiko – wie beispielsweise die Gewährung von Schreibzugriff für einen Agenten auf die Produktionsinfrastruktur – erfordern eine Risikominderung oder eine Begrenzung des Anwendungsbereichs, bevor sie durchgeführt werden.
Bereitstellungen mit geringen Auswirkungen, aber hohem Risiko sollten gänzlich vermieden werden.

Die unverhandelbaren Vorgaben gelten unabhängig von den Umständen: Kundenvertrauen, Sicherheit und Datenschutz, rechtliche und Compliance-Vorgaben sowie betriebliche Stabilität. Die Rolle des CISO in einer KI-beschleunigten Umgebung besteht darin, sicherzustellen, dass das Unternehmen/ die Organisation kluge Risiken eingeht und gleichzeitig leichtsinnige Risiken vermeidet.

Die in diesem Report dargelegten Fakten erzeugen ein Spannungsfeld, für das Unternehmen und Organisationen eine Lösung finden müssen: Die Einführung von KI birgt echte Risiken, aber eine langsame Einführung ist ebenfalls risikobehaftet.

Bei Einführung agentischer KI Schadensradius reduzieren

Sophos hat sieben Maßnahmen vorgeschlagen, die innerhalb der nächsten ein bis sechs Monate umgesetzt werden können, um den potenziellen Schadensradius für Unternehmen und Organisationen zu verringern, die agentische KI einführen möchten:

1. **Agent Sandboxing** setzt eine kontrollierte Grenze um den Agentenprozess. Die meisten Harnesses werden mit einer Art Sandboxing ausgeliefert, das jedoch in der Regel erst aktiviert werden muss. Wählen Sie nach Möglichkeit Remote- oder cloudbasierte Ausführungsumgebungen anstelle von lokalen Sandboxes, um eine bessere Prozesstrennung zu gewährleisten.
2. **Die Isolierung von Zugangsdaten** stellt sicher, dass der Agent niemals direkt Zugriff auf vertrauliche Daten hat: Ein separater Prozess ruft die Zugangsdaten aus einem Tresor ab, fügt sie in den API-Aufruf ein und gibt ausschließlich bereinigte Antworten zurück, wodurch langlebige Zugangsdaten aus den Kontextfenstern des LLM ferngehalten werden.
3. **Versiegelte Tool Endpoints** gehen noch weiter. Der Agent ruft ein Tool namentlich auf, doch ein Broker-Prozess verwaltet die Zugangsdaten, führt den eigentlichen API-Aufruf gemäß einem festgelegten Schema durch und setzt die für jedes Tool geltenden Egress Allowlists durch. Der einzige Hebel des Agenten besteht darin, auszuwählen, welches Tool aufgerufen und welche Parameter übergeben werden sollen.
4. **Egress-Beschränkung und Netzwerküberwachung** behandeln den kompromittierten Agenten als Kanal für die Datenexfiltration: Erkennung vertraulicher Daten im ausgehenden Datenverkehr, Volumenschwellenwerte für ausgehende Anfragen und Webkategorisierung, die Anfragen an nicht kategorisierte oder neu registrierte Domänen blockiert.
5. **Die EDR-Abdeckung** muss sich auch auf das Verhalten der Agent-Hosts erstrecken: Container, kurzlebige VMs und nicht verwaltete Entwicklersysteme sind häufige blinde Flecken.
6. **Human-gated Approval (menschliche Freigabe)** trennt die autonome Ausführungsebene des Agenten von einer von Menschen gesteuerten Kontrollebene: Für irreversible Vorgänge ist ein kryptografischer Vorgang erforderlich und nicht etwa eine einfache „OK"-Schaltfläche, die der Agent simulieren könnte.
7. **Grenzen der Injection-Ausbreitung** adressieren das eskalierende Risiko, wenn sich Injections von Session-Ebene über Agenten-Ebene bis hin zur agentenübergreifenden Ausbreitung bewegen. Behandeln Sie Speicherschreibvorgänge und agentenübergreifende Übergaben als Sicherheitsereignisse.

Als OpenClaw Anfang 2026 an Bedeutung gewann, waren über 30.000 Instanzen im Internet exponiert und Bedrohungsakteure diskutierten bereits darüber, wie sie dessen modulare „Fähigkeiten" für Botnet-Kampagnen einsetzen könnten. Das Red Team von Sophos hat dies direkt getestet: 23 umsetzbare Erkenntnisse, eine Verkürzung der AD-Aufklärung von drei Tagen auf drei Stunden sowie Audit-Details, die manuell nicht innerhalb eines angemessenen Zeitrahmens erzielt werden könnten. Aber das Team verbrachte mehr Zeit damit, Sicherheitsgrenzen aufzubauen, als den Test durchzuführen. Die Lektion: Agentische KI ist leistungsstark genug, um eine Bereitstellung zu rechtfertigen, aber nur, wenn der Sicherheitsrahmen zuerst aufgebaut wird.

Transparenz, Identität und Patching

Wir empfehlen, mit der Transparenz über die Nutzung von KI-Services zu beginnen. Die KI-Exposition, die bei den Vorfällen bei Salesloft/Drift und Vercel dokumentiert wurde, macht dies zum wirkungsvollsten ersten Schritt. Erstellen Sie ein Inventar der KI-Tools in Ihrer Umgebung. Identifizieren Sie, welche davon OAuth-Token, API-Schlüssel oder Dienstkonten enthalten, die mit Unternehmenssystemen verbunden sind. Ermitteln Sie, welche Daten diese passieren und welche Berechtigungen sie besitzen. Was Sie nicht sehen können, können Sie auch nicht steuern.

Im Rahmen der Transparenzmaßnahmen sollte ein KI-Register erstellt werden, in das jedes KI-Tool, jeder KI-Agent, jedes KI-Modell, jede Zugangsberechtigung, auf die Zugriff besteht, jede Sandboxing-Konfiguration sowie alle erwarteten Konnektivitätsanforderungen aufgenommen werden. Ohne dieses Register können Kontrollen wie Isolierung von Zugangsdaten, Egress-Beschränkung und Sandboxing nicht konsistent angewendet werden. Wenden Sie Identitätskontrollen mit der gleichen Strenge an wie bei jeder anderen SaaS-Anwendung: Setzen Sie bedingte Zugriffsrichtlinien durch, verlangen Sie MFA für KI-Dienstkonten, wenden Sie DLP-Kontrollen auf die von KI-Tools genutzten ausgehenden Datenpfade an und rotieren Sie die Zugangsdaten für KI-verbundene Integrationen nach demselben Zeitplan wie bei jedem anderen privilegierten Zugriff. Sowohl die Erkenntnisse von IBM X-Force zu Zugangsdaten als auch der Supply-Chain-Angriff auf die Nx Console bestätigen, dass Zugangsdaten für KI-Services ein Harvesting-Ziel sind.

KI-gestützte Angriffe verkürzen das Zeitfenster zwischen Bekanntwerden und Ausnutzung einer Sicherheitslücke auf wenige Stunden.

Unternehmen und Organisationen, deren Patch-Zyklus noch in Wochen gemessen wird, sehen sich einem zunehmenden Risiko gegenüber. Der in der CISA-Richtlinie BOD 26-04 festgelegte Zeitrahmen von drei Tagen stellt lediglich eine Untergrenze dar, da KI-gestützte Angriffe das Zeitfenster zwischen Bekanntwerden und Ausnutzung auf wenige Stunden verkürzen. Wenn eine kritische, mit dem Internet verbundene Schwachstelle nicht innerhalb weniger Tage gepatcht werden kann, birgt diese Lücke ein messbares Risiko.

Bewerten Sie auf Grundlage von Nachweisen

Statten Sie die KI-Tools, die Ihre Analysten verwenden, mit Messfunktionen aus. Messen Sie den Validierungsaufwand, wie viel Zeit Analysten mit der Überprüfung von KI-Outputs verbringen. Messen Sie die Effizienz der Prompt-Erstellung, wie viele Versuche bis zu nützlichen Ergebnissen nötig sind. Und messen Sie die Vertrauenskalibrierung – ob Analysten dem System zu sehr oder zu wenig vertrauen. Erstellen Sie Testsätze aus echten SOC-Workloads statt generischen Benchmarks und versionieren Sie Ihre Verfahren, damit sie Modell-Upgrades überstehen.

Schützen Sie die KI-Lieferkette

KI bringt Risiken für die Lieferkette mit sich, die über herkömmliche Software-Abhängigkeiten hinausgehen. Modellgewichte, Herkunft der Trainingsdaten, Integrität des MCP-Servers und die Inferenz-Infrastruktur selbst stellen allesamt Vertrauensgrenzen dar. Unternehmen und Organisationen, die Open-Weight-Modelle über Managed-Cloud-Provider einsetzen, müssen vor der Übermittlung von privatem Code oder Daten die genaue Modellversion, die Bereitstellungsregion, die Speicherdauer, die Richtlinien zur Nutzung für Trainingszwecke, den Umfang der Protokollierung sowie die Vertragsbedingungen überprüfen. Für Modelle, die über SaaS-Endpoints mit Standort in der Volksrepublik China bereitgestellt werden, wie beispielweise die API von DeepSeek, gilt eine strengere Einschränkung: Solche Modelle sollten für öffentliche oder synthetische Kapazitätstests verwendet werden, nicht für proprietäre Repositorys oder sensible technische Kontexte.

Fazit

Das Framework der STAC6994-Bedrohungsakteure enthielt Cobalt Strike-Profile, EDR-Umgehungsmodule, Tools für das Harvesting von Zugangsdaten, Dienstprogramme für laterale Bewegungen und einen C2-Kanal, der hinter einer legitimen Cloud-Infrastruktur verborgen war. Jede Komponente war uns bekannt – außer dem Entwicklungsprozess. KI-Agenten entwickelten Evasionstechniken schneller, als es ein menschlicher Entwickler hätte tun können, testeten sie in einem strukturierten Labor an bestimmten Produkten, verwalteten die Ergebnisse über Git in verschiedenen Versionen und übergaben die Ergebnisse an einen Betreiber, der Ransomware einsetzte und Daten stahl.

Sowohl bei den MDR- und IR-Fällen von Sophos als auch bei den Erkenntnissen der CTU, den Analysen der SophosLabs, der KI-Forschung von Sophos und in der Sicherheitsbranche insgesamt beschleunigt KI bekannte Abläufe auf beiden Seiten der Sicherheitsgleichung. Sie hilft Angreifern, bekannte Angriffsketten schneller zu durchlaufen. Sie ermöglicht Verteidigern, in kurzer Zeit eine Triage und Analyse vorzunehmen. Dadurch entstehen neue Risiken im Zusammenhang mit dem Identitäts- und Governance-Layer, der unvermeidlich mit der Einführung von Unternehmens-KI verbunden ist. KI hat keinen grundlegend neuen Angriffstyp hervorgebracht, dem die bestehende Erkennungsarchitektur nicht gewachsen wäre. Das britische AI Security Institute hat über einen Zeitraum von 18 Monaten eine etwa sechsfache Steigerung der autonomen offensiven Fähigkeiten gemessen. Jede Modellgeneration ermöglichte Schritte in mehrstufigen Angriffsszenarien, die frühere Generationen nicht bewältigen konnten. Diese Entwicklung erlaubt keine Selbstgefälligkeit.

Die folgenden drei Punkte sind zentral für eine sichere KI-Einführung:

1. **Die Klassifizierung** muss präzise sein, denn wenn alle KI-Bedrohungen in einen Topf geworfen werden, wird unklar, welche Abwehrmaßnahmen jeweils erforderlich sind.
2. **Die Bewertung** muss streng sein, denn ganz gleich, ob es um Modell-Routing, Alert-Unterdrückung oder die Tiefe der Schlussfolgerungen der Agenten geht – die Antwort sollte auf Messungen beruhen.
3. **Die Transparenz** muss real sein, denn Vertrauen entsteht bei Unternehmen und Organisationen, die ihre Arbeit offenlegen und neben Erfolgen auch aus Fehlern lernen.

Was Sophos in den nächsten 12 Monaten beobachten wird: ob autonome Agenten davon abkommen, bekannte Playbooks auszuführen, und stattdessen in großem Maßstab neue Angriffspfade aufdecken; ob die Schattenwirtschaft produktisierte KI-Angriffstools entwickelt, die die Einstiegshürde ebenso drastisch senken wie „Ransomware-as-a-Service“ die Hürde für Erpressung gesenkt hat; ob die KI-Governance in Unternehmen schnell genug reift, um die durch die aktuelle Einführungswelle entstandene Gefährdung von Identitäten zu beseitigen; und ob sich das Zeitfenster zwischen Ausnutzung einer Schwachstelle und der Bereitstellung eines Patches weiter verkürzt, was Unternehmen und Organisationen, die ihre Reaktionsgeschwindigkeit noch in Wochen messen, zu einer Neubewertung zwingen wird.

Die Zeiten haben sich gewandelt. Verteidiger, die ihr Tempo nicht anpassen, werden feststellen, dass sie weiter zurückliegen, als sie angenommen hatten.

Beitragende

Nash Borges
SVP of
Engineering and AI

Adarsh Kyadige
Senior Manager
AI Research

Ross McKerchar
CISO

Rafe Pilling
Director of Threat
Intelligence
X-Ops Counter Threat Unit

Matt Stec
Director
Sophos AI

Ryan Westman
Senior Manager Threat
Research
X-Ops Insights

Matt Wixey
Senior Threat Researcher
X-Ops Insights

Unser Team

Sophos X-Ops ist eine hochmoderne Cybersecurity-Initiative, die mehr als 1.000 Experten aus verschiedenen spezialisierten Sicherheitsbereichen bei Sophos zusammenbringt, darunter Threat Intelligence, Künstliche Intelligenz, MDR Operations und interne Security Operations.

Die funktionsübergreifende Task Force stärkt die Abwehr von Unternehmen und Organisationen gegen die sich schnell entwickelnden und hochkomplexen Cyberbedrohungen von heute.

Durch die Nutzung der gebündelten Expertise seiner Task Force bietet Sophos X-Ops eine mehrdimensionale Reaktion auf Cyberangriffe und gewährleistet umfassende Schutz-, Erkennungs- und Reaktionsfähigkeiten.

Dieser kollaborative und innovative Ansatz sorgt für eine einzigartige Bedrohungsreaktion und macht Sophos zu einem Spitzenreiter und führenden Cybersecurity-Anbieter.

Sophos X-Ops nutzt die gebündelte Expertise seiner funktionsübergreifenden Task Force. Die Synergie zwischen den funktionsübergreifenden Teams von Sophos X-Ops fördert den Informationsaustausch und ermöglicht es ihnen, sich schnell an neue Erkenntnisse anzupassen – was die Erkennungs- und Reaktionszeiten beschleunigt und gleichzeitig die allgemeinen Schutzfunktionen für Sophos-Kunden stärkt.

SOPHOS

Wenn Sie Ihre Anforderungen an KI-Sicherheit besprechen und erfahren möchten, wie Sophos Ihnen dabei helfen kann, besuchen Sie unsere Website oder sprechen Sie mit einem Berater.

Sales DACH (Deutschland, Österreich, Schweiz)
Tel.: +49 611 58580
E-Mail: sales@sophos.de

2026-07-20 DE (NP) CRE-5820