

KI in der Cyberkriminalität: Warum die alten Erkennungsregeln nicht mehr greifen

Von Thomas Faß · 6. August 2026 · aktualisiert 8. September 2026

Künstliche Intelligenz ist in der Cybersicherheit ein zweischneidiges Schwert: Angreifer nutzen sie für überzeugende Deepfakes, grammatikalisch perfektes Phishing in Masse und anpassungsfähige Schadsoftware; Verteidiger für Anomalieerkennung, automatisierte Reaktion (SOAR/EDR) und stärkere Authentifizierung. Das Ergebnis ist ein Wettrüsten, in dem menschliche Expertise und Governance entscheidend bleiben – KI soll die Verteidigung verstärken, nicht ersetzen.

Das Wichtigste in Kürze

- Bei der Masse ist KI angekommen, bei den gezielten Angriffen noch nicht. Eine Untersuchung von Columbia University, University of Chicago und Barracuda auf der ACM Internet Measurement Conference 2025 wertete 481.558 böartige E-Mails aus und kommt zu dem Schluss: „at least ~51% of spam emails and ~14% of business email compromise attacks in our dataset are generated using LLMs, as of April 2025.“ Die teuersten Angriffe sind bislang überwiegend Handarbeit – der 14-Prozent-Wert ist der, den man im Blick behalten muss.
- KI ist in der Cybersicherheit ein zweischneidiges Schwert: Dieselbe Technologie befeuert Angriffe und stärkt die Verteidigung.
- Offensiv sorgt KI für überzeugende Deepfakes, grammatikalisch perfektes Phishing in Masse und anpassungsfähige Schadsoftware.
- Defensiv liefert KI Anomalieerkennung, automatisierte Reaktion (SOAR/EDR) und stärkere Authentifizierung über Zero Trust und Passkeys.
- Das Ergebnis ist ein Wettrüsten – menschliche Expertise und Governance bleiben entscheidend: KI soll die Verteidigung verstärken, nicht ersetzen.

Einleitung: Die Dualität der KI in der Cyberwelt

Künstliche Intelligenz (KI) hat sich zu einer Schlüsseltechnologie entwickelt, die nahezu alle Lebensbereiche durchdringt – und die Cybersicherheit bildet hier keine Ausnahme. Ihre Fähigkeit, komplexe Muster zu erkennen, riesige Datenmengen zu verarbeiten und autonome Entscheidungen zu treffen, verleiht ihr eine doppelte Rolle: Sie ist sowohl ein mächtiges Werkzeug für Cyberkriminelle als auch ein unverzichtbarer Verbündeter für Verteidiger im Kampf gegen Cyberkriminalität. In einer zunehmend digitalisierten Welt, in der die Geschwindigkeit und Komplexität von Cyberbedrohungen exponentiell wachsen, ist es für technisch versierte Fachkräfte und Entscheidungsträger unerlässlich, die Mechanismen beider Seiten zu verstehen. Nur so können effektive Strategien zur Risikominderung und zur Stärkung der Cyberresilienz entwickelt werden.

Die Cybersicherheitslandschaft wird zunehmend von einem „KI-Rüstungswettlauf“ geprägt. Angreifer nutzen fortschrittliche KI, um ihre Methoden zu verfeinern, während Verteidiger KI einsetzen, um Schritt zu halten und proaktive Schutzmaßnahmen zu entwickeln. Diese dynamische Auseinandersetzung beschleunigt die Angriffsgeschwindigkeiten, wobei Ausbruchszeiten oft unter einer Stunde liegen. Dies verdeutlicht, dass statische Verteidigungsmechanismen nicht mehr ausreichen und eine kontinuierliche Anpassung und Innovation auf beiden Seiten unerlässlich sind. KI fungiert als ein „Force Multiplier“ sowohl für Angreifer als auch für Verteidiger. Für Cyberkriminelle ermöglicht sie eine Skalierung und Personalisierung von Angriffen in einem bisher unerreichten Ausmaß. Für Verteidiger beschleunigt sie die Erkennung und Reaktion, automatisiert Routineaufgaben und entlastet menschliche Analysten. Die Herausforderung besteht darin, diese sich ständig weiterentwickelnde Dynamik zu managen und die eigene Verteidigung kontinuierlich zu stärken.

Die Offensive: Wie generative KI Cyberangriffe befeuert

Deepfakes: Die neue Dimension des Social Engineering und Betrugs

Deepfakes sind hyperrealistische, KI-generierte Medieninhalte wie Videos, Audio oder Bilder, die reale Personen oder Szenen täuschend echt imitieren. Die zugrundeliegenden Technologien umfassen hauptsächlich Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs) und Diffusion Models. Diese Modelle lernen die zugrundeliegenden Muster und Strukturen von Trainingsdaten und nutzen sie, um neue, ähnliche Daten zu erzeugen. Bei GANs konkurrieren beispielsweise ein Generator und ein Diskriminator, um immer überzeugendere Fälschungen zu produzieren.

Deepfakes werden zunehmend für Social Engineering und Betrug eingesetzt, indem sie die Glaubwürdigkeit von Identitäten und Botschaften drastisch erhöhen. Kriminelle imitieren Führungskräfte, um Mitarbeiter zu betrügen und hohe Geldsummen zu überweisen. Die Entwicklung lässt sich an vier Fällen ablesen. 2019 verlor ein ungenanntes britisches Energieunternehmen 243.000 US-Dollar durch Stimmklonung. 2020 verlor ein Bankmanager in den Vereinigten Arabischen Emiraten rund 35 Millionen US-Dollar durch einen Phishing-Angriff mit Deepfake-Sprachklon. 2024 traf es das britische Ingenieurbüro Arup: In der Hongkonger Niederlassung wurde ein Mitarbeiter in einen Videoanruf mit Deepfakes des Finanzvorstands und weiterer Kollegen gezogen und überwies daraufhin 25,6 Millionen US-Dollar (200 Millionen Hongkong-Dollar) – nicht auf einen Schlag, sondern in fünfzehn Einzeltransaktionen über eine Woche. Ein Unternehmenssprecher bestätigte im Mai 2024: „We can confirm that fake voices and images were used“ – und ebenso: „none of our internal systems were compromised“. Der zweite Satz ist der lehrreichere. Kein System war kompromittiert. Es gab eine Woche Zeit und fünfzehn Gelegenheiten, zurückzufragen, und keine wurde genutzt. Auch prominente Unternehmen wie Ferrari und WPP waren Ziele solcher Angriffe, die jedoch dank wachsamer Mitarbeiter vereitelt werden konnten. Im Frühjahr 2025 wurden zudem italienische Wirtschaftsführer, darunter Giorgio Armani, Opfer einer Welle von Deepfake-Angriffen, bei denen Kriminelle den italienischen Verteidigungsminister Guido Crosetto imitierten und Opfer dazu brachten, eine Million Euro zu überweisen. Neben der CEO-Impersonation werden manipulierte Videos von Prominenten wie Elon Musk verwendet, um Opfer zur Investition in betrügerische Krypto-Plattformen zu verleiten. Die Entwicklung von Live-Deepfakes ermöglicht zudem die Manipulation von Gesichtern und Stimmen in Echtzeit-Videoanrufen, was die Erkennung extrem erschwert und eine neue Ebene der Raffinesse in Betrugsversuche einführt.

Die zunehmende Verfügbarkeit und Kosteneffizienz von Deepfake-Tools führt zu einer Demokratisierung der Angriffswerkzeuge, was die Angriffsfläche erheblich erweitert und die Verteidigung vor neue Herausforderungen stellt. Die Kosten für Deepfake-Tools können „so wenig wie 5 bis 10 US-Dollar pro Monat“ betragen, und einige sind sogar kostenlos. Zudem wird oft nur „eine Minute Audio“ benötigt, um ein überzeugendes Ergebnis zu erzielen. Diese Entwicklung senkt die Eintrittsbarriere für Cyberkriminelle erheblich. Wie stark sich das ausgewirkt hat, hat Sumsb gemessen: Deepfake-Vorfälle haben sich von 2022 auf 2023 weltweit verzehnfacht, in Nordamerika lag der Anstieg bei 1.740 Prozent. Zu den frühen Fällen zählt ein Bankmanager in den Vereinigten Arabischen Emiraten, der 2020 rund 35 Millionen US-Dollar durch einen Phishing-Angriff mit Deepfake-Sprachklon verlor. Die Folge ist eine Zunahme des Volumens und der Qualität von Deepfake-Angriffen, da auch Angreifer mit geringeren technischen Fähigkeiten und Budgets hochwirksame Deepfake-Angriffe durchführen können.

Der Erfolg von Deepfake-Angriffen beruht nicht nur auf der technologischen Perfektion, sondern maßgeblich auf der Ausnutzung menschlicher Emotionen und kognitiver Verzerrungen. Deepfakes sind „visuell und psychologisch am wirkungsvollsten“, da sie „Vertrauen aufbauen“ und „ein Gefühl der Panik oder Dringlichkeit erzeugen“ können. Beispielsweise nutzten Kriminelle den „Patriotismus“ italienischer Geschäftsleute aus, um sie zu manipulieren. Dies zeigt, dass selbst bei technischer Skepsis die emotionale Manipulation entscheidend sein kann. Die fortschrittliche KI ermöglicht es Angreifern, Social Engineering-Taktiken effektiver zu gestalten, indem sie menschliches Vertrauen und emotionale Reaktionen gezielt ausnutzen. Dies unterstreicht, dass technische Abwehrmaßnahmen durch Schulungen zur Sensibilisierung und kritischem Denken ergänzt werden müssen, um die menschliche Komponente als primäres Angriffsziel zu schützen.

KI-gesteuerte Phishing-Kampagnen: Personalisierung und Skalierung

Phishing-Angriffe haben sich von einfachen E-Mail-Betrügereien zu hochraffinierten, KI-gestützten Kampagnen entwickelt, die schwer zu identifizieren sind. Große Sprachmodelle (LLMs) wie GPT können grammatikalisch perfekte, kontextuell relevante und kulturell nuancierte Nachrichten generieren. KI ermöglicht Cyberkriminellen, Tausende personalisierter Phishing-E-Mails innerhalb von Sekunden zu generieren. Dies umfasst die Anpassung von Sprache, Ton und Stil an regionale Kommunikationsnormen. KI wird eingesetzt, um E-Mail-Erkennungssysteme zu umgehen und bösartige Nachrichten glaubwürdiger zu machen. Dies geschieht durch subtile Textvariationen und A/B-Tests von Formulierungen, um die Wirksamkeit zu maximieren.

KI-gesteuerte Phishing-Angriffe beschränken sich nicht mehr nur auf E-Mails. Sie umfassen eine Kombination aus E-Mail-Kommunikation, die den Ton und Stil einer Organisation perfekt nachahmt, Sprachsynthese, die die Stimme eines CEOs klonen kann für telefonische Nachfassaktionen, Videomanipulation zur Erstellung überzeugender Deepfake-Meetings und Echtzeit-Chat-Antworten, die eine konsistente Imitation über verschiedene Kanäle hinweg aufrechterhalten. Darüber hinaus ermöglicht KI-gestützte Datenanalyse Angreifern, riesige Datensätze aus sozialen Medien, öffentlichen Aufzeichnungen und kompromittierten Datenbanken in beispielloser Geschwindigkeit zu sammeln und zu analysieren. Dies erleichtert hochgradig zielgerichtete Spear-Phishing-Kampagnen, die auf spezifische Personen oder Organisationen zugeschnitten sind.

Die Fähigkeit von LLMs, „fehlerfreie“ und „kontextuell relevante“ Phishing-E-Mails zu erstellen, bedeutet, dass herkömmliche Spamfilter, die auf Grammatikfehler oder offensichtliche Muster abzielen, zunehmend

ineffektiv werden. Eine Untersuchung von Columbia University, University of Chicago und Barracuda, veröffentlicht auf der ACM Internet Measurement Conference 2025, hat 481.558 bösartige E-Mails aus der Zeit vor und nach der Veröffentlichung von ChatGPT ausgewertet – nach der Bereinigung 239.145 Spam- und 242.413 BEC-Mails. Der zentrale Befund im Wortlaut: „at least ~51% of spam emails and ~14% of business email compromise attacks in our dataset are generated using LLMs, as of April 2025.“ Für den Alltag heißt das zweierlei: Bei der Masse – Werbung, Betrugsversuche in Serie, alles, was ungezielt im Postfach landet – ist die KI angekommen, gut die Hälfte im April 2025. Bei Business Email Compromise, also den gezielten Angriffen auf einzelne Personen mit vorbereiteter Legende, waren es 14 Prozent. Die teuersten Angriffe sind bislang überwiegend Handarbeit. Das ist kein Grund zur Entwarnung, sondern zur Einordnung: Der Anteil bei den gezielten Angriffen hat den längeren Weg vor sich. Die KI-generierten Nachrichten wiesen dabei eine „höhere Formalität, weniger Grammatikfehler und größere sprachliche Raffinesse“ auf. Dies zwingt die Verteidigung dazu, sich von signaturbasierten Filtern zu komplexeren NLP-Modellen und Verhaltensanalysen zu bewegen. Die sprachliche Raffinesse von KI-generierten Phishing-E-Mails verstärkt die Notwendigkeit von KI-gestützten Verteidigungsmechanismen auf einer höheren Abstraktionsebene. Unternehmen, die sich ausschließlich auf signaturbasierte Antiviren-Lösungen verlassen, sind einem erhöhten Risiko ausgesetzt.

Die Kombination von Daten-Scraping und LLM-Fähigkeiten ermöglicht eine „Hyper-Personalisierung“ von Phishing-Angriffen. Dies geht über das bloße Einfügen eines Namens hinaus und zielt auf individuelle Verhaltensweisen und Interessen ab. KI kann „Benutzerverhalten und digitale Kommunikationsmuster analysieren“, um „hochgradig zielgerichtete Spear-Phishing-Angriffe“ zu erstellen. Dies bedeutet, dass Angreifer nicht nur technische Schwachstellen, sondern auch psychologische Experten der Opfer ausnutzen. Die Implikation ist, dass Schulungen zur Cybersicherheit nicht nur technische Indikatoren, sondern auch psychologische Manipulationstaktiken abdecken müssen, da die menschliche Komponente zum primären Ziel wird.

Die Nutzung mehrerer Kommunikationskanäle (E-Mail, Telefon, Video, Chat) in einem einzigen Phishing-Angriff erhöht die Komplexität der Erkennung erheblich. Wenn ein Angreifer „nahtlos zwischen E-Mail, Sprache und Video wechseln kann, während eine konsistente Nachahmung beibehalten wird“, werden traditionelle, isolierte Sicherheitsmaßnahmen ineffektiv. Dies erfordert, dass Verteidigungssysteme eine integrierte, kanalübergreifende Überwachung und Korrelation von Bedrohungsdaten nutzen, um solche komplexen Angriffe zu erkennen und zu stoppen. Silo-basierte Sicherheitslösungen sind in diesem Szenario unzureichend.

Adaptive Malware und „Agentware“: Autonome Bedrohungen

Adaptive Malware ist eine fortschrittliche Form bösartiger Software, die KI und maschinelles Lernen (ML) nutzt, um sich kontinuierlich weiterzuentwickeln und so die Erkennung und Eliminierung erheblich zu erschweren. Ihre Hauptmerkmale umfassen selbstmodifizierenden Code (Polymorphismus), der seine Struktur kontinuierlich ändert, um Antiviren-Signaturen zu umgehen. Sie verwendet dynamische Malware-Payloads, die bösartige Skripte an spezifische Ziele und Umgebungen anpassen, und nutzt KI-gesteuerte Tarnung, um legitime Anwendungen nachzuahmen und mit dem Netzwerkverkehr zu verschmelzen. Adaptive Malware ist zudem in der Lage, in Echtzeit aus ihrer Umgebung zu lernen und ihr Verhalten anzupassen, um Sicherheitsmaßnahmen zu umgehen. Sie kann „eigenständig denken“ und ihr Verhalten unabhängig ändern, um bestehende Cybersicherheitsmaßnahmen zu umgehen. Im Gegensatz zu statischer

Malware, die vordefinierten Anweisungen folgt, lernt adaptive Malware aus fehlgeschlagenen Angriffen, passt Angriffe an und verbreitet sich autonom.

Das Konzept der „Agentware“ oder Agentic AI in Cyberangriffen bezieht sich auf autonome KI-Systeme, die in der Lage sind, Entscheidungen zu treffen und Aktionen ohne direkte menschliche Eingabe auszuführen. Sie fungieren als „Agenten“, die ihre Umgebung wahrnehmen, Informationen verarbeiten und Aktionen zur Erreichung ihrer Ziele ausführen können. Zu ihren autonomen Fähigkeiten gehören das Scannen von Netzwerken, das Identifizieren von Schwachstellen, das Starten gezielter Angriffe und das Modifizieren des eigenen Codes zur Umgehung von Erkennungssystemen ohne menschliches Eingreifen. Sie können aus früheren Angriffen lernen und ihre Strategien verfeinern.

Der Einsatz autonomer KI in Cyberangriffen wirft erhebliche ethische Bedenken auf. Agentic AIs sind anfällig für „adversarial exploits“, bei denen Hacker Eingaben oder Trainingsdaten manipulieren, um das Verhalten der KI zu verzerren. Dies kann dazu führen, dass die KI Bedrohungen übersieht oder falsche Elemente markiert. Wie alle KIs können autonome Agenten unter Halluzinationen, Bias und unberechenbarem Verhalten leiden, was durch ihre Autonomie verstärkt werden kann. Die Fähigkeit dieser Systeme, in Echtzeit und ohne menschliche Aufsicht Entscheidungen zu treffen, kann zu unbeabsichtigten Konsequenzen führen, wie der falschen Identifizierung von Bedrohungen oder der Eskalation von Cyberangriffen. Die Zurechnung von Verantwortung bei Fehlern oder Kollateralschäden wird problematisch. KI-Systeme könnten zudem mehrdeutige Situationen unbeabsichtigt falsch interpretieren und Konflikte eskalieren lassen, indem sie auf wahrgenommene Bedrohungen mit unverhältnismäßiger Gewalt reagieren.

Die Kernunterscheidung zwischen traditioneller und adaptiver Malware liegt in der Verschiebung von statischen Anweisungen zu kontinuierlichem Lernen und Echtzeit-Anpassung. Dies macht signaturbasierte Erkennung zunehmend ineffektiv. Polymorphe Malware kann ihren Code kontinuierlich mutieren, was es „schwierig für traditionelle signaturbasierte Erkennungsmethoden macht, die Malware zu erkennen und zu blockieren“. Die KI-gesteuerte Dynamik in Angriffen erfordert eine KI-gesteuerte Dynamik in der Verteidigung, die auf Verhaltensanalyse und Anomalieerkennung statt auf bekannten Mustern basiert. Unternehmen, die sich ausschließlich auf signaturbasierte Antiviren-Lösungen verlassen, sind einem erhöhten Risiko ausgesetzt.

Das Konzept der „Agentware“ führt die Möglichkeit ein, dass Cyberangriffe von KI-Systemen autonom und ohne direkte menschliche Aufsicht durchgeführt werden. Dies stellt eine signifikante Eskalation dar, da Angriffe mit Maschinengeschwindigkeit ablaufen könnten. Agentic AI kann „autonom Netzwerke scannen, Schwachstellen identifizieren und gezielte Angriffe starten, ohne menschliches Eingreifen“. Dies führt zu einer „zunehmenden Komplexität und Geschwindigkeit moderner Bedrohungen“. Die Implikation ist ein zukünftiges Szenario, in dem Cyberangriffe und -verteidigungen primär zwischen KI-Systemen stattfinden könnten, was die menschliche Reaktionsfähigkeit überfordern würde. Dies erfordert eine Neudefinition der Rolle des Menschen in der Cybersicherheit.

Mehrere Bedenken betonen die ethischen Bedenken hinsichtlich autonomer Systeme, insbesondere in Bezug auf Verantwortlichkeit, Transparenz und Bias. Es wird argumentiert, dass die „Zurechnung von Verantwortung“ bei Fehlern autonomer Systeme „problematisch“ ist, da die Entscheidungsprozesse oft „Black Boxes“ sind. Zudem können Trainingsdaten „eingebettete Verzerrungen“ enthalten, die zu

diskriminierenden Ergebnissen führen. Die technische Entwicklung ohne gleichzeitige Entwicklung robuster ethischer und rechtlicher Rahmenwerke kann zu unkontrollierbaren und potenziell schädlichen Ergebnissen führen. Dies erfordert einen interdisziplinären Ansatz, der Rechtswissenschaftler, Ethiker und Technologen zusammenbringt.

Die Defensive: Wie KI die Cyberversicherung stärkt

Künstliche Intelligenz (KI) spielt eine immer wichtigere Rolle in der Cybersicherheit, insbesondere bei der Anomalieerkennung, prädiktiven Bedrohungsanalyse und der Automatisierung von Abwehrmaßnahmen. Sie ist entscheidend, um mit der rasanten Geschwindigkeit und Komplexität moderner Cyberangriffe Schritt zu halten.

Anomalieerkennung und prädiktive Bedrohungsanalyse

KI-gesteuerte Verhaltensanalyse zielt darauf ab, „normale“ Aktivitäten in Systemen und Netzwerken zu verstehen und Abweichungen davon zu erkennen. Sie passt sich kontinuierlich an verändertes Benutzerverhalten und Systemkonfigurationen an. KI-Systeme nutzen dafür verschiedene Techniken des maschinellen Lernens, einschließlich unüberwachtem Lernen, das besonders effektiv bei der Erkennung unbekannter Bedrohungen (Zero-Days) ist. Anwendungsfälle sind die Identifizierung von Insider-Bedrohungen oder kompromittierten Konten durch ungewöhnliche Anmeldeuster, Datenzugriffe oder Befehlszeilenaktivitäten. Auch eine Frühwarnung vor Datenexfiltration ist möglich.

Im Bereich der prädiktiven Bedrohungsanalyse und Vulnerability Assessment analysiert KI kontinuierlich Bedrohungsfeeds aus globalen Quellen, verarbeitet natürliche Sprachbeschreibungen von Bedrohungen und korreliert scheinbar unzusammenhängende Sicherheitsereignisse, um potenzielle Angriffskampagnen in ihren frühen Phasen zu identifizieren. Sie kann zudem Code-Muster, Systemkonfigurationen und Architekturdesigns analysieren, um potenzielle Schwachstellen vorherzusagen, die noch nicht öffentlich bekannt sind, einschließlich der Analyse von Software-Abhängigkeiten und Drittanbieter-Komponenten.

Die Vorteile sind eine verbesserte Genauigkeit bei der Bedrohungserkennung, Echtzeit-Analyse und -Reaktion, Reduzierung von Fehlalarmen, proaktive Bedrohungsjagd sowie minimale Auswirkungen auf die Benutzererfahrung. Herausforderungen bleiben jedoch die Datenüberflutung, die Schwierigkeit bei der Definition von „Normalverhalten“, hohe Raten von Fehlalarmen, die Anpassung an sich ständig entwickelnde Bedrohungen und der Mangel an Fachkräften.

Die Beschleunigung von Cyberangriffen macht manuelle menschliche Reaktionen zunehmend unzureichend. KI-gesteuerte Anomalieerkennung ist eine direkte Antwort darauf, indem sie Abweichungen vom Normalverhalten identifiziert, auch bei unbekannten Bedrohungen. Dies ist ein fundamentaler Paradigmenwechsel von der reaktiven Abwehr bekannter Angriffe zur proaktiven Identifizierung neuartiger Bedrohungen. Die zunehmende Raffinesse von Angriffen verstärkt die Notwendigkeit von prädiktiven und verhaltensbasierten Verteidigungsmethoden. Unternehmen müssen kontinuierlich in die Weiterentwicklung ihrer KI-Modelle investieren, um mit der sich ständig ändernden Bedrohungslandschaft Schritt zu halten.

Angesichts des Fachkräftemangels im Bereich Cybersicherheit und der schieren Datenmenge, die analysiert werden muss, ist die manuelle Bedrohungserkennung nicht mehr skalierbar. KI-Systeme können

riesige Datenmengen in Echtzeit analysieren und Fehlalarme reduzieren. KI-gesteuerte Lösungen ermöglichen es menschlichen Sicherheitsexperten, sich auf strategische Sicherheitsinitiativen zu konzentrieren und wichtigere Sicherheitsaufgaben zu übernehmen. Dies ist eine direkte Antwort auf den Fachkräftemangel, indem die Effizienz und der Fokus der bestehenden Teams maximiert werden. KI verbessert nicht nur die Sicherheit, sondern kann auch die Arbeitszufriedenheit der Analysten erhöhen und die Bindung von Talenten fördern, indem sie repetitive und ermüdende Aufgaben automatisiert.

Automatisierung von Abwehrmaßnahmen und Incident Response

SOAR-Plattformen (Security Orchestration, Automation, and Response) koordinieren und automatisieren Sicherheitsaufgaben und -prozesse über verschiedene Sicherheitstools und -systeme hinweg. Dies umfasst die Automatisierung von Bedrohungserkennung, -untersuchung und -eindämmung sowie das Management des gesamten Incident-Response-Lebenszyklus. KI in SOAR lernt Verhaltensmuster, sagt Bedrohungen voraus und automatisiert Reaktionen. KI-Agenten werden in Security Operations Centern (SOCs) immer häufiger eingesetzt, um Routineaufgaben zu übernehmen und menschliche Analysten zu unterstützen. Dazu gehören Aufgaben wie die Triage von Alarmen, die Durchsuchung von Logs nach Bedrohungen und die Automatisierung grundlegender Incident-Responses.

Automatisierte Reaktionen umfassen das Blockieren, Isolieren und Beheben von Bedrohungen. Endpoint Detection and Response (EDR)-Tools können kompromittierte Endpunkte automatisch erkennen und isolieren, bösartige Prozesse beenden oder bösartige Dateien blockieren, um Angriffe frühzeitig einzudämmen. Risikobasierte Authentifizierungssysteme, wie sie beispielsweise in Office 365 zum Einsatz kommen, können riskante Anmeldungen erkennen (z. B. von unbekannten Standorten) und automatische Aktionen wie das Widerrufen von Benutzersitzungen oder das Zurücksetzen von Passwörtern auslösen. Tools können Phishing-E-Mails automatisch erkennen und unter Quarantäne stellen, bevor sie den Posteingang des Benutzers erreichen.

Die Vorteile der Automatisierung sind vielfältig: Sie beschleunigt die Reaktionszeiten erheblich, minimiert Schäden durch schnelle Eindämmung, reduziert menschliche Fehler, steigert die Effizienz durch Entlastung der Sicherheitsteams von repetitiven Aufgaben und gewährleistet Konsistenz und Compliance bei der Ausführung von Incident-Response-Prozessen. Herausforderungen der Automatisierung sind jedoch Fehlalarme, die Geschäftsabläufe stören können, wenn sie legitime Aktivitäten blockieren. Eine kontinuierliche Feinabstimmung der Erkennungsregeln ist hier entscheidend. Eine übermäßige Abhängigkeit von Automatisierung kann ebenfalls problematisch sein; Sicherheitsteams sollten eine „Set-it-and-forget-it“-Mentalität vermeiden, da menschliche Aufsicht und Urteilsvermögen unerlässlich bleiben. Zudem erfordert die nahtlose Interaktion zwischen verschiedenen Sicherheitstools sorgfältige Planung und kontinuierliche Tests.

Die Beschleunigung von Cyberangriffen, mit Ausbruchszeiten oft unter einer Stunde, macht manuelle menschliche Reaktionen zunehmend unzureichend. SOAR und KI-gesteuerte Automatisierung sind eine direkte Antwort darauf. SOAR-Plattformen sind darauf ausgelegt, „Reaktionszeiten zu beschleunigen“ und „Schäden zu minimieren“. Die exponentielle Zunahme der Angriffsgeschwindigkeit erzwingt eine Automatisierung der Abwehrmaßnahmen auf Maschinengeschwindigkeitsebene. Unternehmen, die ihre Incident-Response-Prozesse nicht automatisieren, müssen mit erheblich höheren Schäden und längeren Ausfallzeiten rechnen.

KI-gesteuerte Automatisierung im SOC geht über die bloße Effizienzsteigerung hinaus. Sie ermöglicht eine strategische Allokation knapper menschlicher Ressourcen. KI-Agenten können „Aufgaben mit geringerem Risiko automatisieren“, wodurch Teams „sich auf Bedrohungen mit hoher Priorität konzentrieren“ können. Dies ist eine direkte Antwort auf den Fachkräftemangel in der Cybersicherheit. Dies führt zu einer Maximierung des Nutzens menschlicher Expertise, indem repetitive, zeitaufwändige Aufgaben an die KI delegiert werden. Dies führt zu einer effizienteren und effektiveren Sicherheitslage und kann die Arbeitszufriedenheit der Analysten verbessern.

Obwohl Automatisierung entscheidend ist, warnen mehrere Stimmen vor „False Positives“ und „Over-Reliance on Automation“ und betonen, dass KI „Menschen nicht ersetzen kann“. Automatisierte Abhilfemaßnahmen sind nicht perfekt und können „Geschäftsabläufe stören“, wenn sie legitime Aktivitäten fälschlicherweise blockieren. Dies unterstreicht die Notwendigkeit menschlicher Aufsicht für komplexe Entscheidungen und zur Minderung unbeabsichtigter Folgen. Erfolgreiche Implementierungen von Automatisierung in der Cybersicherheit erfordern ein sorgfältiges Design der Zusammenarbeit zwischen Mensch und KI, anstatt eine vollständige Autonomie der KI anzustreben. Das Vertrauen in die KI muss durch Transparenz und Erklärbarkeit (Explainable AI, XAI) gestärkt werden, um die Akzeptanz und Wirksamkeit zu gewährleisten.

Stärkung der Authentifizierung und Zugriffsverwaltung durch Zero Trust und Passkeys

Die [Zero Trust](#) Architecture (ZTA) ist ein Sicherheitsrahmen, der auf dem Prinzip „Niemals vertrauen, immer verifizieren“ basiert. Sie eliminiert das implizite Vertrauen, das in traditionellen perimeterbasierten Modellen angenommen wird, und erfordert eine kontinuierliche Verifizierung jeder Zugriffsanfrage, unabhängig vom Standort des Benutzers oder Geräts. Die Kernprinzipien der ZTA, nach NIST 800-207, umfassen die kontinuierliche Verifizierung und Authentifizierung jeder Entität (Benutzer, Gerät, Anwendung), idealerweise bei jeder Anfrage. Dies beinhaltet starke Multi-Faktor-Authentifizierung (MFA) und die Bewertung kontextueller Signale. Das Prinzip der geringsten Rechte (Least Privilege Access) gewährt Benutzern und Geräten nur den minimal erforderlichen Zugriff, um ihre Aufgaben zu erfüllen. Dies minimiert den „Blast Radius“ eines potenziellen Verstoßes. Mikrosegmentierung unterteilt das Netzwerk in kleinere, isolierte Segmente, um die laterale Bewegung von Angreifern zu begrenzen und den Angriffsradius zu reduzieren. Schließlich ist die kontinuierliche Überwachung und Analyse des Benutzerverhaltens, des Netzwerkverkehrs und der Systemaktivitäten zur Erkennung von Anomalien und potenziellen Bedrohungen in Echtzeit ein zentraler Bestandteil.

Die ZTA bietet zahlreiche Vorteile, darunter eine reduzierte Angriffsfläche, verbesserte Sicherheit, erhöhte Compliance mit Vorschriften wie GDPR, HIPAA und PCI DSS, gesteigerte Agilität für neue Technologien und Cloud-Dienste sowie bessere Sichtbarkeit und Kontrolle. Die Implementierung der ZTA bringt jedoch auch Herausforderungen mit sich. Der kulturelle Wandel, der mit der Abkehr von traditionellen Arbeitsweisen und der Angst vor Kontrollverlust einhergeht, kann Widerstand hervorrufen. Lösungen hierfür sind klare Kommunikation der Vorteile, Führungskräfte als Vorbilder und agile Change-Management-Ansätze. Die Integration von Legacy-Systemen ist eine weitere Hürde, da ältere Systeme oft nicht für ZTA konzipiert sind. Hier können Middleware, API-Gateways und Identity-Proxy-Dienste die Lücke überbrücken. Schließlich stellt der Mangel an Fachkräften mit ZTA-Expertise eine Herausforderung dar, der durch umfassende Schulungen und Upskilling der Teams begegnet werden kann. Zu den Schlüsselkomponenten der ZTA gehören Identity and Access Management (IAM) als Grundlage, Secure

Access Service Edge (SASE), Data Loss Prevention (DLP), Security Information and Event Management (SIEM), Unified Endpoint Management (UEM) und Policy-based Enforcement Points.

Passkeys sind FIDO-Authentifizierungsnachweise, die auf FIDO-Standards (FIDO2, WebAuthn, CTAP) basieren und Passwörter durch kryptografische Schlüsselpaare (Public-Key-Kryptographie) ersetzen. Die Authentifizierung erfolgt über den Entsperrmechanismus des Geräts (Biometrie, PIN). Die Vorteile von Passkeys sind vielfältig: Sie sind hochgradig Phishing-resistent, da sie an die Identität einer Website oder App gebunden sind und nicht auf betrügerischen Websites verwendet werden können. Die Eliminierung von Passwort-Wiederverwendung ist ein weiterer Vorteil, da jeder Passkey einzigartig für eine Website/App ist. Die Benutzerfreundlichkeit wird durch schnellere Anmeldungen, die Eliminierung der Notwendigkeit, komplexe Passwörter zu merken oder zu verwalten, und die nahtlose Synchronisierung über Geräte hinweg erheblich verbessert. Zudem reduzieren Passkeys den IT-Overhead, da administrative Aufgaben und Support-Anfragen für Passwort-Resets entfallen. Die Branchenakzeptanz ist hoch, wobei große Technologieunternehmen wie Apple, Google und Microsoft Passkeys aktiv integrieren. Microsoft ist diesen Weg konsequent gegangen: Seit Juni 2025 lassen sich in der Authenticator-App keine Kennwörter mehr hinzufügen, seit Juli 2025 ist das automatische Ausfüllen abgeschaltet, und seit Mitte August 2025 greift die App nicht mehr auf gespeicherte Zugangsdaten zu; erreichbar bleiben diese über Microsoft Edge. Neue Microsoft-Konten sind seit Mai 2025 standardmäßig passwortlos. Herausforderungen bei der Einführung von Passkeys umfassen die Kontowiederherstellung, die oft noch an Passwörter oder Benutzer-IDs gebunden ist, die Geräteabhängigkeit, da Verlust oder Ausfall des Geräts den Zugriff erschweren kann, sowie die Kompatibilität zwischen verschiedenen Plattformen und Geräten. Potenzial für Anbieterbindung (Vendor Lock-in) besteht ebenfalls, wenn Passkeys nicht einfach zwischen verschiedenen Plattformen übertragen werden können. Schließlich kann mangelnde Benutzerschulung und -akzeptanz die Einführung behindern.

Im Vergleich zu traditionellen 2FA/MFA-Methoden eliminieren Passkeys die Notwendigkeit eines Passworts vollständig, während 2FA eine zweite Verifizierungsstufe zu einem Passwort hinzufügt. Traditionelle 2FA-Methoden wie SMS-OTPs sind anfällig für SIM-Swapping und Abfangen. Passkeys sind gegen solche Angriffe immun, da sie keine Geheimnisse übermitteln. Passkeys erfüllen die Anforderungen an Multi-Faktor-Authentifizierung (MFA) in einem einzigen, nahtlosen Schritt, indem sie Gerätebesitz („etwas, das man hat“) und Biometrie („etwas, das man ist“) kombinieren.

Der Paradigmenwechsel von perimeterbasierter zu identitätszentrierter Sicherheit durch ZTA ist eine grundlegende Veränderung, die durch Remote Work, Cloud-Einführung und raffinierte Bedrohungen vorangetrieben wird, die herkömmliche Perimetermodelle umgehen. In einer Welt, in der Benutzer von überall und mit beliebigen Geräten auf Ressourcen zugreifen, ist der physische Netzwerkperimeter bedeutungslos geworden. ZTA verlagert den Fokus auf „kontinuierliche Authentifizierung und Autorisierung jeder Zugriffsanfrage, unabhängig von ihrem Ursprung oder Standort“. Die zunehmende Verteilung von Arbeitskräften und IT-Ressourcen macht die traditionellen Sicherheitsmodelle ineffektiv und erfordert eine identitätsbasierte, kontextsensitive Sicherheitsstrategie. Unternehmen müssen ihre gesamte Sicherheitsarchitektur neu gestalten, um die Identität des Benutzers und den Kontext des Zugriffs als primäre Vertrauensfaktoren zu behandeln.

ZTA wird nicht nur als Sicherheitsmaßnahme, sondern als strategischer Enabler für Cloud-basierte und Remote-Arbeitsumgebungen dargestellt. ZTA „erhöht die Agilität“ bei der Einführung neuer Technologien

und Cloud-Dienste und ist „ideal für Organisationen, die auf Remote Work angewiesen sind“. Es ermöglicht Unternehmen, die Vorteile von Multi-Cloud-Umgebungen und flexiblen Arbeitsmodellen sicher zu nutzen. ZTA bietet die notwendige Sicherheit, um die inhärenten Risiken verteilter IT-Infrastrukturen zu mindern, wodurch Unternehmen Innovationen vorantreiben können, ohne die Sicherheit zu kompromittieren. ZTA ist somit nicht nur eine IT-Sicherheitsentscheidung, sondern eine strategische Geschäftsentscheidung, die die Wettbewerbsfähigkeit und Anpassungsfähigkeit eines Unternehmens direkt beeinflusst.

Passkeys werden als direkte Antwort auf die Schwachstellen traditioneller Authentifizierungsmethoden präsentiert, die oft auf menschlichen Fehlern (schwache Passwörter, Wiederverwendung, Phishing-Anfälligkeit) basieren. Durch den Einsatz von Public-Key-Kryptographie und Biometrie auf dem Gerät eliminieren Passkeys die Notwendigkeit, Passwörter zu merken oder einzugeben, und machen Phishing-Angriffe weitgehend unwirksam. Die „massive Akzeptanz“ durch große Tech-Unternehmen wie Apple, Google und Microsoft signalisiert einen bedeutenden Wandel in der Branche. Die Schwachstellen passwortbasierter Systeme und traditioneller 2FA-Methoden treiben die Entwicklung und Akzeptanz von Passkeys voran. Dies führt zu einer Zukunft, in der Authentifizierung nicht nur sicherer, sondern auch deutlich benutzerfreundlicher wird, was die Akzeptanz von Sicherheitsmaßnahmen durch Endbenutzer erheblich verbessern kann. Gleichzeitig bleibt die Herausforderung der Migration von Legacy-Systemen und der Benutzerschulung bestehen.

Phishing-Resistenz ist ein entscheidender Faktor, bei dem Passkeys mit „Hoch“ bewertet werden, da der private Schlüssel das Gerät nie verlässt und die Authentifizierung an die Domain gebunden ist. Passwörter, auch mit 2FA, und SMS-OTPs weisen hier eine geringere bis mittlere Resistenz auf, da sie anfällig für Diebstahl oder Abfangen sind. In puncto Benutzerfreundlichkeit überzeugen Passkeys mit einer „sehr hohen“ Bewertung, da der Anmeldevorgang nahtlos durch Gerätesperren erfolgt. Passwörter mit 2FA und SMS-OTPs erfordern zusätzliche Schritte und bieten daher nur mittlere Benutzerfreundlichkeit. Die Angriffsfläche ist bei Passkeys „gering“, da keine Passwörter auf dem Server gespeichert werden. Bei passwortbasierten Systemen ist die Angriffsfläche „mittel bis hoch“, während SMS-OTPs ebenfalls eine mittlere Angriffsfläche aufweisen. Ein wesentlicher Vorteil von Passkeys ist, dass Passwörter nicht auf dem Server gespeichert werden müssen. Dies ist bei traditionellen Passwörtern der Fall, während es bei SMS-OTPs nicht zutreffend ist, da es sich um einen zweiten Faktor handelt. Der Schutz vor Passwort-Wiederverwendung ist bei Passkeys gewährleistet, da für jeden Dienst einzigartige Schlüsselpaare verwendet werden. Bei Passwörtern ist dieser Schutz nicht gegeben, wenn Benutzer Passwörter wiederverwenden. SMS-OTPs schützen, da sie Einmalcodes sind. Die Abhängigkeit von einem Mobilfunknetz für OTPs ist ein Nachteil von SMS-OTPs, während Passkeys diese Abhängigkeit nicht haben. In Bezug auf die Kosten für den Anbieter sind Passkeys tendenziell „gering“, da keine Transaktionsgebühren anfallen. SMS-OTPs können hier „mittel bis hoch“ sein. Schließlich bieten Passkeys eine „hohe“ Cross-Device Kompatibilität durch Cloud-Synchronisation, während traditionelle Methoden hier variabler sind.

Der KI-Cybersecurity-Rüstungswettlauf: Ausblick und Fazit

Der Wettlauf zwischen Cyberkriminellen und Cyberverteidigern, befeuert durch KI, ist ein kontinuierlicher Innovationszyklus. Angreifer entwickeln ständig neue, raffiniertere Methoden, während Verteidiger mit adaptiven KI-Lösungen reagieren müssen. Diese Dynamik bedeutet, dass Cybersicherheit keine einmalige Implementierung, sondern ein fortlaufender Prozess ist.

Die Effektivität von defensiver KI hängt maßgeblich von der Qualität der Daten ab, mit denen sie trainiert wird. Hochwertige, kontextbezogene und gut annotierte Datensätze sind entscheidend, um effektive KI-Modelle zu trainieren, die zwischen gutartigem und böartigem Verhalten unterscheiden können. Ohne eine robuste Datengovernance und die Fähigkeit, Daten aus heterogenen Quellen zu konsolidieren und zu verarbeiten, wird die Sicherheits-KI ihre Mission nicht erfüllen können.

Trotz der zunehmenden Autonomie von KI-Systemen bleibt die menschliche Komponente unverzichtbar. KI soll menschliche Experten augmentieren, nicht ersetzen. Es ist entscheidend, eine „Set-it-and-forget-it“-Mentalität zu vermeiden. Menschliche Analysten sind weiterhin für komplexe Entscheidungen, die Bewertung von Fehlalarmen und die strategische Weiterentwicklung der Sicherheitssysteme zuständig. Die „Awareness and Governance Gap“ stellt hier eine erhebliche Schwachstelle dar.

Die zunehmende Autonomie von KI in der Cybersicherheit wirft tiefgreifende ethische Fragen auf, insbesondere in Bezug auf Verantwortlichkeit, Transparenz, Bias und den Schutz der Privatsphäre. Die Entwicklung robuster ethischer Rahmenwerke und klarer gesetzlicher Regelungen ist unerlässlich, um sicherzustellen, dass KI verantwortungsvoll eingesetzt wird und unbeabsichtigte Schäden vermieden werden. Konzepte wie Explainable AI (XAI) gewinnen an Bedeutung, um die „Black-Box“-Natur von KI-Entscheidungen zu adressieren.

Empfehlungen für Unternehmen: Strategien zur Stärkung der Cyberresilienz in einer KI-dominierten Landschaft

Um in diesem dynamischen KI-Cybersecurity-Rüstungswettlauf zu bestehen und die Cyberresilienz zu stärken, sollten Unternehmen folgende strategische Maßnahmen ergreifen:

- **Investition in KI-gesteuerte Verteidigungstools:** Unternehmen sollten proaktiv in KI-basierte Anomalieerkennung, prädiktive Analyse und automatisierte Reaktion (SOAR/EDR) investieren, um Bedrohungen in Echtzeit zu identifizieren und zu neutralisieren. Die zunehmende Raffinesse und Skalierung von KI-Angriffen treibt die Investitionen in defensive KI voran. Es ist eine strategische Notwendigkeit, um im „Rüstungswettlauf“ zu bestehen.
- **Implementierung einer Zero Trust Architektur (ZTA):** Die Abkehr vom perimeterbasierten Sicherheitsmodell hin zu einem „Niemals vertrauen, immer verifizieren“-Ansatz ist unerlässlich. Dies umfasst die kontinuierliche Verifizierung von Identitäten, das Prinzip der geringsten Rechte und die Mikrosegmentierung von Netzwerken. ZTA ermöglicht es, die Sicherheit in verteilten und dynamischen Cloud-Umgebungen zu gewährleisten. Die zunehmende Verteilung von Arbeitskräften und IT-Ressourcen macht traditionelle Sicherheitsmodelle ineffektiv und erfordert eine identitätsbasierte, kontextsensitive Strategie.
- **Einführung passwortloser Authentifizierung (Passkeys):** Passkeys bieten eine überlegene Sicherheit gegenüber Passwörtern und traditionellen 2FA-Methoden, indem sie Phishing-Angriffe wirksam abwehren und die Benutzerfreundlichkeit verbessern. Die Integration sollte strategisch geplant werden, um Kompatibilitätsprobleme zu minimieren. Passkeys sind eine direkte Antwort auf die Schwachstellen traditioneller Authentifizierungsmethoden, die oft auf menschlichen Fehlern basieren.

- **Priorisierung von kontinuierlichem Lernen und Upskilling:** Angesichts des schnellen Wandels der KI-Technologien müssen Unternehmen kontinuierlich in die Weiterbildung ihrer IT- und Sicherheitsteams investieren. Regelmäßige Schulungsprogramme, die aktuelle Bedrohungsmuster und den Umgang mit Deepfakes und KI-Phishing simulieren, sind entscheidend. Die Komplexität der KI-Bedrohungen und -Verteidigungen erfordert eine gleichzeitige Entwicklung von Technologie, Prozessen und menschlichen Fähigkeiten.
- **Förderung einer Sicherheitskultur:** Sicherheit muss als gemeinsame Verantwortung aller Mitarbeiter verstanden und in alle Geschäftsprozesse integriert werden (DevSecOps). Dies erfordert eine offene Kommunikation und Zusammenarbeit zwischen den Abteilungen.
- **Implementierung robuster Governance- und Compliance-Frameworks:** Insbesondere für den Einsatz von KI-Systemen und den Umgang mit Daten sind klare Richtlinien und automatisierte Compliance-Prüfungen unerlässlich.
- **Strategische Nutzung von Multi-Cloud- und Hybrid-Cloud-Umgebungen:** Diese Architekturen bieten erhöhte Resilienz, Flexibilität und Kosteneffizienz, erfordern jedoch ein sorgfältiges Management der Komplexität, insbesondere in Bezug auf Sicherheit und Governance.

Die zunehmende Raffinesse und Skalierung von KI-Angriffen treibt die Investitionen in defensive KI voran. Eine große Mehrheit der Sicherheitsverantwortlichen hält KI-basierte Automatisierung für wichtig, um neuen Bedrohungen einen Schritt voraus zu sein. Dies zeigt, dass Investitionen in defensive KI nicht optional, sondern eine strategische Notwendigkeit sind, um im „Rüstungswettlauf“ zu bestehen. Die offensive Nutzung von KI durch Cyberkriminelle erfordert eine entsprechende defensive Reaktion, um die Cyberresilienz aufrechtzuerhalten. Unternehmen, die hier nicht mithalten, sind einem erheblich höheren Risiko ausgesetzt, Opfer von Cyberangriffen zu werden.

Trotz des technologischen Fortschritts betonen die Analysen immer wieder die Bedeutung menschlicher Faktoren. Der „Awareness and Governance Gap“ ist eine große Schwachstelle, da nur wenige Organisationen umfassende KI-Richtlinien oder Mitarbeiterschulungen haben. Dies bedeutet, dass selbst die fortschrittlichsten KI-Verteidigungssysteme durch menschliche Fehler oder mangelndes Verständnis untergraben werden können. Die Komplexität der KI-Bedrohungen und -Verteidigungen erfordert eine gleichzeitige Entwicklung von Technologie, Prozessen und menschlichen Fähigkeiten. Eine umfassende Cybersicherheitsstrategie muss Investitionen in Technologie, kontinuierliche Schulung der Mitarbeiter und die Förderung einer starken Sicherheitskultur umfassen.

Die Diskussion über ZTA und die Rolle von KI in der Reduzierung von Erkennungs- und Reaktionszeiten zeigt, dass Cybersicherheit nicht mehr nur eine IT-Funktion ist, sondern direkt zur Geschäftssagilität und zum Geschäftsergebnis beiträgt. ZTA wird als „Sicherheit für digitale Unternehmen“ beschrieben, die „Geschäftssagilität“ und „kontinuierlich sich entwickelnde Bedrohungen“ adressiert. Es ermöglicht Unternehmen, die „digitale Transformation“ voranzutreiben. In der heutigen digitalen Wirtschaft ist eine robuste Cybersicherheit eine Voraussetzung für Innovation, Wachstum und Wettbewerbsfähigkeit. Cybersicherheitsentscheidungen müssen auf höchster Führungsebene getroffen und mit den übergeordneten Geschäftszielen abgestimmt werden, um den maximalen Wert zu erzielen und Risiken effektiv zu managen.

Ihr nächster Schritt

Die Regel „achten Sie auf Rechtschreibfehler“ hat ausgedient. Was sie ersetzt, ist kein Werkzeug, sondern ein Prozess: ein zweiter Kanal für jede Zahlungsfreigabe, der nicht derselbe ist wie der, über den die Anweisung kam. Wir richten das mit Ihnen ein — mit Senior-Experten aus unserer [Cybersecurity-Expertise](#), deutschsprachig und gesteuert aus Deutschland.

Häufige Fragen

Wie nutzen Angreifer KI konkret?

Vor allem für Täuschung im großen Stil: Deepfakes von Stimmen und Videos für Betrug (etwa gefälschte Chef-Anweisungen), grammatikalisch perfekte und hyper-personalisierte Phishing-Mails in Sekunden, dazu anpassungsfähige Schadsoftware, die ihre Erkennung erschwert. KI senkt die Hürde und erhöht die Wirksamkeit von Angriffen.

Wie hilft KI der Verteidigung?

Sie erkennt Muster, die Menschen übersehen: Anomalieerkennung findet ungewöhnliches Verhalten und Zero-Day-nahe Bedrohungen, SOAR- und EDR-Systeme automatisieren Erkennung, Untersuchung und Eindämmung, und risikobasierte Authentifizierung reagiert automatisch auf verdächtige Anmeldungen. So sinken Reaktionszeiten und Fehlalarme.

Was bleibt trotz KI unverzichtbar?

Der Mensch und die Governance. KI soll augmentieren, nicht ersetzen — die Qualität der Trainingsdaten und nachvollziehbare Entscheidungen (Explainable AI) sind kritisch. Die größte Schwachstelle ist oft die Awareness- und Governance-Lücke, nicht die Technik selbst.

Was sind KI-generierte Deepfakes und warum sind sie im Social Engineering so gefährlich?

Deepfakes sind mit KI erzeugte, täuschend echte Bild-, Audio- oder Videofälschungen. Im Social Engineering machen sie Betrug überzeugender: Eine gefälschte Stimme oder ein gefälschtes Video einer Führungskraft kann Mitarbeitende zu Überweisungen oder Datenfreigaben verleiten. Weil die Fälschung menschliche Wahrnehmung gezielt ausnutzt, greifen rein technische Filter allein nicht — Prozesse mit Rückkanal-Verifikation werden wichtiger.

Wie verändert generative KI Phishing-Angriffe?

Generative KI erzeugt grammatikalisch fehlerfreie, individuell zugeschnittene Nachrichten in großer Menge. Die früheren Warnsignale — holprige Sprache, offensichtliche Fehler — verschwinden, und Kampagnen lassen sich stark personalisieren und skalieren. Erkennung verschiebt sich damit von Textmerkmalen hin zu Verhaltens- und Kontextanalyse.

Was ist adaptive Malware bzw. "Agentware"?

Adaptive Malware verändert ihr Verhalten dynamisch, um Erkennung zu entgehen; als "Agentware" agiert sie zunehmend eigenständig und trifft eigene Entscheidungen im Zielsystem. Das erschwert signaturbasierte Abwehr und verlangt verhaltensbasierte Erkennung und schnelle, teils automatisierte Reaktion.

Welche Rolle spielen Zero Trust und Passkeys in der KI-gestützten Verteidigung?

Zero Trust beschränkt Zugriffe auf das minimal Nötige und prüft sie kontinuierlich, sodass ein kompromittiertes Konto weniger Schaden anrichtet. Passkeys ersetzen Passwörter durch phishing-resistente Authentifizierung – genau dort, wo KI-Phishing ansetzt. Zusammen verkleinern sie die Angriffsfläche, die KI-gestützte Angriffe ausnutzen.

Thematische Vertiefungen

[Cybersecurity-Expertise](#)

[Ransomware 2.0](#)

[Zero Trust](#)

Quellen

- Sumsb: Global Deepfake Incidents Surge Tenfold from 2022 to 2023. Belegt die Verzehnfachung der Deepfake-Vorfälle von 2022 auf 2023 und den Anstieg um 1.740 % in Nordamerika. – <https://sumsub.com/newsroom/sumsub-research-global-deepfake-incidents-surge-tenfold-from-2022-to-2023/>
- ENISA: Threat Landscape (EU-Bedrohungslage) – <https://www.enisa.europa.eu/topics/cyber-threats/threat-landscape>
- Europäische Kommission: AI Act (Regulierungsrahmen) – <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
- Wei Hao u. a.: Do Spammers Dream of Electric Sheep? Characterizing the Prevalence of LLM-Generated Malicious Emails. ACM Internet Measurement Conference 2025, 12 Seiten. Datensatz: 481.558 bösartige E-Mails, nach Bereinigung 239.145 Spam- und 242.413 BEC-Mails. Belegt wörtlich: „at least ~51% of spam emails and ~14% of business email compromise attacks in our dataset are generated using LLMs, as of April 2025.“ – <https://www.cs.columbia.edu/~junfeng/papers/ai-email-imc25.pdf>
- Columbia Engineering: AI now powers over half of spam emails – Meldung der Hochschule zur Studie, ergänzend – <https://www.ee.columbia.edu/news/ai-now-powers-over-half-spam-emails-columbia-engineering-research-finds>
- Fortune: Arup revealed as victim of deepfake scam, 17. Mai 2024. Belegt 25,6 Mio. US-Dollar (200 Mio. Hongkong-Dollar) in fünfzehn Transaktionen über eine Woche, die betroffene Hongkonger Niederlassung und die beiden Unternehmenszitate. – <https://fortune.com/europe/2024/05/17/arup-deepfake-fraud-scam-victim-hong-kong-25-million-cfo/>
- Microsoft Support: Änderungen am automatischen Ausfüllen von Microsoft Authenticator. Belegt die drei Stufen Juni, Juli und Mitte August 2025 sowie den Verbleib der Kennwörter in Microsoft Edge. – <https://support.microsoft.com/de-de/account-billing/%C3%A4nderungen-am-automatischen-ausf%C3%BCllen-von-microsoft-authenticator-09fd75df-dc04-4477-9619-811510805ab6>