Schilde hoch

Eine Anleitung für die Abwehr im Zeitalter KI-gestützter Angriffe

Zusammenfassung

Anfang 2026 begannen Branchenführer aufgrund potenzieller Sicherheitsrisiken den Zugriff auf ihre fortschrittlichsten KI-Modelle einzuschränken. Cisco war dabei führend, arbeitete mit Anthropic an deren Mythos Preview-Modell zusammen und erhielt Zugriff auf GPT-5.5-Cyber von OpenAI. Diese Partnerschaften sind Teil einer umfassenderen, andauernden Initiative, um unsere Abwehrmechanismen gegen hochmoderne KI zu testen. Dabei sind wir uns bewusst, dass sich unsere Zusammenarbeit mit neuen aufkommenden Modellen weiterentwickeln wird.

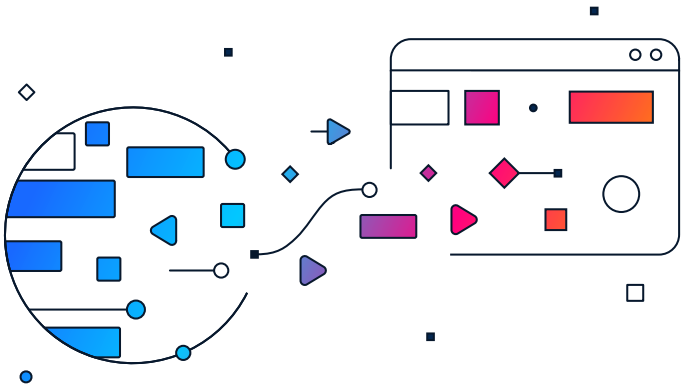
Unsere Erfahrungen mit diesen Modellen haben dazu geführt, dass wir unsere zukünftige Bedrohungsmodellierung an die Angreifer des KI-Zeitalters angepasst haben. Dies hat wiederum unsere Abwehrmechanismen verändert und wir haben eine Reihe von Empfehlungen zur Abwehr für unsere Kunden entwickelt. Auch wenn die Funktionen hochmoderner KI-Modelle derzeit noch nicht allgemein verfügbar sind, gehen wir davon aus, dass diese und weitere Funktionen es im Zuge der Weiterentwicklung der KI-Technologie früher oder später sein werden.

In diesem Whitepaper erfahren Sie, was Cisco bisher zu KI-gestützten Funktionen beobachtet hat und wie die neue Bedrohungslandschaft unserer Meinung nach aussehen wird. Unabhängig davon, ob diese Modelle von Angreifern eingesetzt, von Forschenden genutzt oder als Agents in Ihrer eigenen Umgebung bereitgestellt werden – die Auswirkungen auf die Sicherheit sind erheblich. Wir erläutern, was wir basierend auf diesen neuen Erkenntnissen implementiert haben, und legen unsere Empfehlungen für Kunden dar.

Die Angriffsfläche wird sich verändern – und das teils drastisch. Verteidiger müssen sich die Zeit nehmen, zu verstehen, wie die neue Normalität aussehen wird, und zu bewerten, welche Änderungen in ihrer Umgebung nötig sind, um die Sicherheit zu gewährleisten. Cisco steht ihnen bei dieser Transformation als Partner zur Seite.

Die Rolle von KI bei aktuellen Cybersicherheits-Events

Schon vor Mythos und GPT-5.5 haben böswillige Akteure KI in ihre Angriffsabläufe integriert. Anfang 2024 haben Microsoft und OpenAI jeweils eine Untersuchung zur böswilligen Verwendung von großen Sprachmodellen (LLMs) veröffentlicht. Microsoft gab damals an, dass „noch keine besonders ungewöhnlichen oder einzigartigen KI-gestützten Angriffs- oder Missbrauchstechniken beobachtet wurden“. Die Dokumentation zeigt zumeist APT-Akteure (Advanced Persistent Threat), die LLMs nutzen, um Informationen in Bereichen wie Satelliten-Kommunikation, Übersetzung technischer Dokumente, Unterstützung bei der Programmierung und Planung von Social-Engineering-Angriffen zu sammeln.



Seit diesem Bericht waren die Akteure aber nicht untätig. In einer Veröffentlichung zu TA547 vermutet Proofpoint, dass die Akteure LLM zur Generierung von PowerShell-Skripten verwendet haben. Cisco hat ebenfalls ein modulares Framework namens VoidLink identifiziert – ein Tool, das über umfangreiche Funktionen wie rollenbasierte Zugriffskontrolle, Peer-to-Peer- und Dead-Letter-Queue-Routing sowie eine Implant-Management-Funktion verfügt. In der Codebasis wurden eine Reihe von Indikatoren gefunden, die darauf hindeuten, dass die Lösung wahrscheinlich mit Unterstützung eines LLM entwickelt wurde.

Insbesondere das Social Engineering hat vom Einsatz von KI profitiert. Es wird von zahlreichen Fällen berichtet, in denen Akteure LLMs zur Verbesserung von E-Mail-Ködern nutzen. Einige Akteure gehen jedoch noch einen Schritt weiter: Mandiant berichtet über den potenziellen Einsatz von KI-Videotools durch UNC1069 zur Erstellung eines Deepfake-Videos, das vermeintlich vom CEO des Zielunternehmens stammt.

Dies sind mit Sicherheit nicht alle beobachteten Fälle von KI-Nutzung durch Angreifer. Sie stehen aber exemplarisch für die Art von Funktionen, von deren böswilliger Nutzung wir bei der Erörterung von Möglichkeiten zur Verteidigung gegen KI-gestützte Akteure ausgingen. Die Funktionen hochmoderner KI-Modelle verändern die Art und Weise, wie wir die Bedrohungslandschaft bewerten.

Die neue KI-Bedrohungslandschaft

Basierend auf unserer Erfahrung mit hochmodernen KI-Modellen ändern wir bei Cisco die Art und Weise, wie wir unsere Angreifer modellieren. Die Funktionen dieser Modelle würden – wären sie allgemein verfügbar – wahrscheinlich zu einer drastischen Verringerung der erforderlichen Kompetenzen für bestimmte Arten von Exploit-Aktivitäten führen. Die Folge wäre eine größere Anzahl von Schwachstellen und damit verbundenen Exploits sowie eine größere Anzahl von Akteuren, die diese Schwachstellen wahrscheinlich ausnutzen würden.

Während der potenzielle Umfang dieser Veränderung alle Verteidiger betreffen würde, wären diejenigen besonders anfällig, die End-of-Life- oder End-of-Support-Geräte oder -Software nutzen. Werden bei diesen Produkten Schwachstellen aufgedeckt, sind Verteidiger besonders anfällig für Angriffe und verfügen über keinerlei effektive Korrekturmaßnahmen.

Diese fortschrittlichen Modelle steigern die Fähigkeiten von Akteuren jeder Art deutlich. Gewöhnliche Akteure werden weiterhin weitgehend opportunistisch handeln, jedoch mit der Möglichkeit, bisher durch ihre Ressourcen beschränkte Aktivitäten zu skalieren. Akteure auf höheren Ebenen mit spezifischeren Zielen können Schwachstellen im Ziel-Technologie-Stack leichter aufdecken. Dies wird die Zeit zwischen Exploit-Versuchen auf bevorzugte Ziele reduzieren.

Werden diese Modelle als Grundlage für KI-Agents verwendet, stellen sie für Angreifer eine neue Möglichkeit dar, wenn sie den Agenten kompromittieren können. Hochmoderne KI-Modelle sollten in streng kontrollierten Sandbox-Umgebungen mit starker Eindämmung betrieben werden. Wie von Cisco beobachtet und von Anthropic im technischen Bericht zu Security-Funktionen von Mythos Preview bestätigt, zeigt das Modell eine hohe Performance bei der Baseline-Ausrichtung, weist jedoch seltene Fehler mit hohem Schweregrad auf, die wie folgt kategorisiert werden:

- Zielorientiertes, strategisches Denken
- Partiale Entkoppelung zwischen interner Kognition und Ausgabe
- Optimierung in Bezug auf implizite oder falsch angegebene Ziele
- „Situationsbewusstsein“, das das Verhalten beeinflusst

Diese Verhaltensweisen stehen im Einklang mit einem aufkommenden Agent-ähnlichen kognitiven Profil statt einem rein reaktiven Sprachmodell. Dieses „situationsbewusste“ Verhalten entspricht nicht dem, was wir von einem Standard-LLM erwarten würden. Herkömmliche LLMs werden als Next-Token-Predictor betrachtet, die mit lokalen Mustern in Texten arbeiten, nicht als Systeme, die ein kohärentes Modell ihrer Umgebung, ihres Kontextes oder ihrer Rolle innerhalb eines breiteren Prozesses pflegen. Ein LLM „weiß“ nicht, ob es evaluiert, bereitgestellt, eingeschränkt

oder beobachtet wird. Es reagiert einfach nur entsprechend den statistischen Korrelationen in der Eingabe. *Das von Anthropic beobachtete und bestätigte Verhalten impliziert jedoch, dass das Modell latente Repräsentationen des Interaktionskontexts selbst bildet (z. B. Bewertungseinstellungen, Einschränkungen oder Benutzerabsichten erkennt) und sein Verhalten entsprechend anpasst.*

Dabei handelt es sich um eine Verlagerung von einer rein reaktiven Vervollständigung von Mustern zu einer kontextsensitiven, selbstreflexiven Argumentation, bei der das Modell Aspekte der Situation implizit über die unmittelbare Eingabeaufforderung hinaus verfolgt. Derartige Funktionen ähneln Elementen der agentischen Kognition (einschließlich Umgebungsmodellierung und bedingter Strategieauswahl), die über das erwartete Verhalten eines Systems hinausgehen, das nur auf die Vorhersage von Text trainiert wird, und daher eine qualitativ andere und komplexere Klasse von Modellverhalten repräsentieren.

Neue Modelle ermöglichen es Angreifern, über ihre Fähigkeiten hinaus zu agieren. Angreifer können schneller reagieren und neue Zero-Day-Schwachstellen aufdecken – selbst in komplexen Stacks. Die Art und Weise, wie wir Abwehrmaßnahmen priorisieren und konzipieren, muss sich ändern, um dieser Bedrohung zu begegnen.

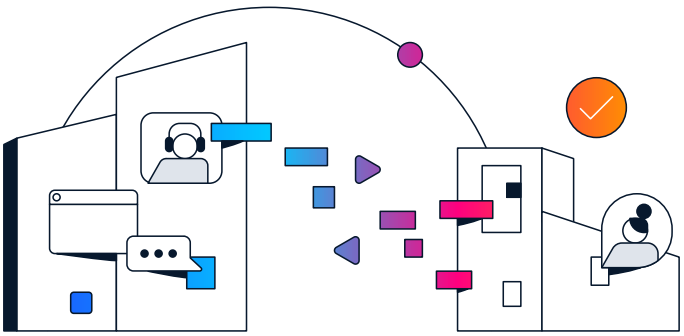
Wie Cisco sich anpasst, um unsere Produkte zu schützen

Cisco stellt sich der Herausforderung der Cyberabwehr im KI-Zeitalter, indem es diese fortschrittlichen KI-Modelle verwendet, um Schwachstellen in einer Geschwindigkeit und einer Skalierung zu finden und zu beheben, die zuvor undenkbar waren. Gleichzeitig beschleunigt Cisco die Entwicklung von Sicherheitsprodukten, die KI-gestützte Angreifer abwehren können. Neben der Erkennung von Schwachstellen und der Produktentwicklung entwickeln wir auch die Art und Weise weiter, wie wir Software erstellen und validieren.

Dazu gehört es, unsere Bedrohungsmodelle zu aktualisieren, um KI-gestützte Angreifer zu berücksichtigen, Szenarien für das KI-Zeitalter in unsere Red-Team-Übungen zu integrieren und letztlich traditionelle Taktiken, Techniken und Verfahren (TTPs) hinter uns zu lassen, um unsere Produkte gegen die tatsächlichen Funktionen dieser Modelle zu testen.

Da KI-Code-Agents zunehmend in Softwareentwicklungs-Workflows integriert werden, muss sichergestellt werden, dass diese Agents standardmäßig sicheren Code erzeugen. Cisco hat kürzlich Project CodeGuard an die Coalition for Secure AI (CoSAI) gespendet. Project CodeGuard bietet ein modellunabhängiges Open-Source-Sicherheits-Framework, das Secure-by-Default-Verfahren direkt in die Workflows von KI-Code-Agents integriert. CodeGuard stellt Sicherheitsfähigkeiten und -regeln bereit, die KI-Agents beibringen, häufige Schwachstellen bei der Codegenerierung und -überprüfung zu vermeiden. Cisco empfiehlt Unternehmen, Frameworks wie CodeGuard einzuführen, um sicherzustellen, dass die KI-Unterstützung, die das Schreiben von Code beschleunigt, nicht unbeabsichtigt auch die Schwachstellen mit sich bringt, die von KI-gestützten Angreifern ausgenutzt werden.

Um unsere Abwehrfunktionen weiter auszubauen, hat Cisco die Foundry Security Spec veröffentlicht, ein Open-Source-Testpaket, das Unternehmen beim Aufbau eines agentenbasierten Systems zur Sicherheitsbewertung unterstützen kann. Diese bewährte Spezifikation ermöglicht es Teams, ein KI-System aufzubauen, das ihre individuellen Umgebungs- und Sicherheitsanforderungen erfüllt. Durch Bereitstellung eines gemeinsamen Frameworks für die Sicherheitsbewertung unterstützen wir die Branche beim Aufbau zuverlässigerer, sichererer Systeme mit Maschinengeschwindigkeit.



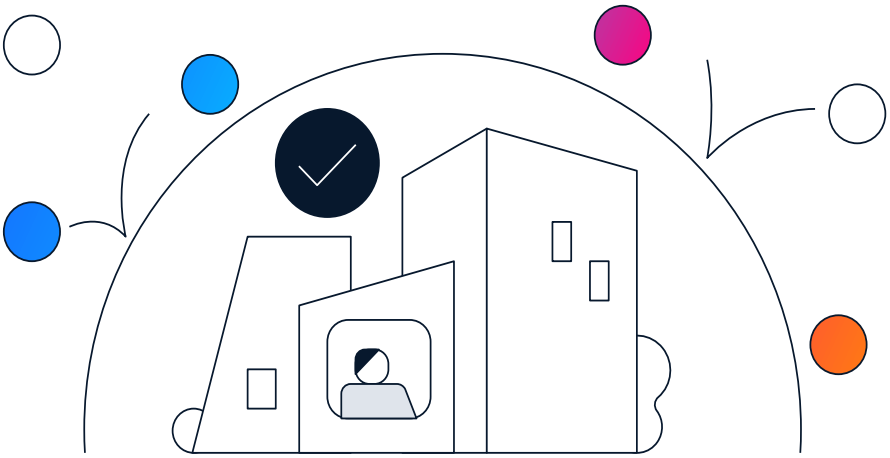
Parallel dazu setzen wir diese Funktionen im Rahmen unserer Initiative für ausfallsichere Infrastruktur ein, in deren Fokus Secure-by-Default- und Secure-by-Design-Prinzipien, proaktive Härtung der Infrastruktur, striktes Patching und Lebenszyklusverwaltung sowie die systematische Abschaffung unsicherer Funktionen und Protokolle in Cisco Produkten stehen.

Dies umfasst die Optimierung von Standardkonfigurationen, die Verbesserung von Protokollierung und Monitoring für eine umfassendere Sicherheitstelemetrie sowie die Modernisierung der Geräteauthentifizierung mit stärkeren Protokollen und Verschlüsselungen. All dies zielt darauf ab, die Angriffsfläche zu reduzieren und Kunden zu helfen, zukünftige Bedrohungen frühzeitig zu erkennen und zu verhindern. Diese Maßnahmen untermauern das Engagement von Cisco, nicht nur auf neue Bedrohungen aus dem KI-Zeitalter zu reagieren, sondern ihnen auch einen Schritt zuvorzukommen und unseren Kunden dabei zu helfen, ein noch ausfallsicheres digitales Fundament aufzubauen.

Unsere frühzeitige Verwendung von hochmodernen KI-Modellen zeigt, dass das alte Modell „ein CVE pro Schwachstelle“ an seine Grenzen kommt. Da die automatisierte Erkennung zu einem exponentiellen Anstieg der identifizierten Bugs führt, verlangsamt die Behandlung jeder kleinen Schwachstelle als individueller Offenlegungsbericht das Sicherheits-Ecosystem und verzögert aktiv die Softwareaktualisierung. Unser Ziel ist es, Kunden mit umsetzbaren Informationen zu unterstützen, statt sie mit Daten zu überschwemmen. Mithilfe eines konsolidierten Offenlegungsmodells, bei dem schwerwiegende Schwachstellen priorisiert und kleinere Korrekturen in die standardmäßigen Release-Zyklen eingebunden werden, können wir Patching-Entscheidungen beschleunigen. Dieser optimierte Ansatz verweigert Angreifern die detaillierten Roadmaps, die sie benötigen, um KI gegen unsere Infrastruktur einzusetzen.

Um uns gegen moderne Bedrohungen zu wappnen, müssen wir Maßnahmen vor der Verwaltung priorisieren. Der herkömmliche Ansatz, jedem geringfügigen Problem einen einzelnen CVE zuzuweisen, führt zu einem Mehraufwand, der Upgrades verlangsamt und Sicherheitsteams überlastet. Wir sind überzeugt, dass sich die Offenlegung zukünftig auf das Ergebnis konzentrieren muss: Kunden zu schnellen Behebungen und Upgrades zu verhelfen. Wir benötigen ein starkes CVE-Programm für die Branche, das für dieses neue Ausmaß an Erkennung und Offenlegung von Sicherheitslücken skalierbar ist.

Diese Maßnahmen untermauern das Engagement von Cisco, nicht nur auf neue Bedrohungen aus dem KI-Zeitalter zu reagieren, sondern ihnen auch einen Schritt zuvorzukommen und unseren Kunden dabei zu helfen, ein noch ausfallsicheres digitales Fundament aufzubauen.



Schutz für unser Unternehmen

Wir wenden diese Prinzipien auch auf unsere eigene Unternehmensumgebung an. Die nachfolgend aufgeführten Empfehlungen sind nicht theoretisch. Sie spiegeln den gleichen Ansatz wider, den Cisco intern zur Abwehr von Bedrohungen des KI-Zeitalters verfolgt. Von der Beschleunigung von Patch-Zyklen und der Beseitigung von End-of-Life-Systemen über die Bereitstellung von KI-gestütztem Threat Hunting bis hin zur Durchsetzung des Least-Privilege-Prinzips für KI-Agents setzen wir diesen Leitfaden auch in unserer eigenen Infrastruktur um.

Unsere Empfehlungen

Um effektiv auf die zunehmenden Fähigkeiten fortschrittlicher KI-Modelle reagieren zu können, **müssen Unternehmen einen ausgewogenen Ansatz verfolgen, der grundlegende Sicherheitspraktiken stärkt und gleichzeitig ihre Abwehrarchitektur modernisiert**. Obwohl sich die Bedrohungslandschaft rasant weiterentwickelt, nutzen viele erfolgreiche Angriffe nach wie vor bekannte Schwachstellen aus. Die Stärkung zentraler Kontrollen ist nach wie vor eine der wirksamsten Maßnahmen, die Security-Führungskräfte ergreifen können.

Unternehmen sollten **grundlegende Maßnahmen wie Phishing-sichere Authentifizierung, starke Identitätsprüfung, Zugriff nach dem Least-Privilege-Prinzip (einschließlich KI-Agents) und Zero-Trust-Architekturen priorisieren**. Konsistentes Patch-Management, umfassende Ressourcentransparenz und diszipliniertes Konfigurationsmanagement sind unerlässlich, um ausnutzbare Schwachstellen zu reduzieren. Diese Kontrollen bilden die Grundlage für Widerstandsfähigkeit und sind entscheidend, wenn es darum geht, den Umfang und die Verbreitung von herkömmlichen und KI-basierten Angriffen zu begrenzen. In vielen Fällen führt die Verbesserung der Umsetzung dieser Grundlagen zu einer schnelleren Risikominderung als die Bereitstellung neuer Technologien.

Gleichzeitig müssen Unternehmen rigorose Maßnahmen ergreifen, um strukturelle Risiken zu beseitigen. **Alle Geräte und Software, die nicht gepatcht, aktualisiert oder unterstützt werden können, müssen systematisch entfernt und durch moderne Plattformen ersetzt werden.** Moderne Systeme integrieren fortschrittliche Schutzmaßnahmen wie Arbeitsspeicher-Sicherheitsmechanismen und Exploit-Abwehr, die die Nutzung von Schwachstellen für Angriffe erheblich erschweren. Selbst wenn Schwachstellen vorhanden sind, bremsen diese Schutzmaßnahmen Angreifer aus und verringern die Wahrscheinlichkeit einer erfolgreichen Ausnutzung. Der Aufbau von Umgebungen, die flexibel, kontinuierlich aktualisierbar und auf schnelles Patching ausgelegt sind, ist heutzutage essenziell – insbesondere für mit dem Internet verbundene Services, bei denen zwischen Offenlegung und Massen-Exploit nur sehr wenig Zeit bleibt.

Aber die Stärkung der Grundlagen und die Modernisierung der Infrastruktur allein reichen nicht aus. Die Geschwindigkeit KI-basierter Angriffe wird die Zeit zwischen der Erkennung von Schwachstellen und deren Ausnutzung auf Minuten oder Sekunden reduzieren. Herkömmliche Modelle, die ausschließlich auf Erkennung und Reaktion basieren, reichen alleine nicht mehr aus. **Verteidiger müssen ihr Betriebsmodell weiterentwickeln, um der Geschwindigkeit, dem Umfang und der Anpassungsfähigkeit von Bedrohungen des KI-Zeitalters gerecht zu werden.** Dazu gehören Investitionen in Erkennung mit Maschinengeschwindigkeit, automatisierte Vorselektierung und Eindämmung sowie kontinuierliches Monitoring von Identitäts- und Datenaktivitäten. Dies reduziert die Abhängigkeit von manuellen Eingriffen und ermöglicht schnellere und konsistentere Reaktionen auf mit hoher Zuverlässigkeit erkannte Bedrohungen.

Diese Entwicklung **erfordert auch eine Umstellung auf integrierte aktive Abwehrmechanismen**. Statt sich ausschließlich auf die Erfassung von Telemetriedaten und Analysen nach Events zu verlassen, sollten Unternehmen Schutzmaßnahmen direkt in Workloads, auf Geräten und im Dateiverkehrspfad platzieren, damit Sicherheitskontrollen in Echtzeit reagieren können. Beispiele hierfür sind Inline-Durchsetzungsmechanismen, Laufzeitschutz mit Technologien wie eBPF für Low-Level-Transparenz und -Kontrolle sowie unabhängig aktualisierbare Exploit Schields, die auf neue Bedrohungen reagieren können, ohne dass vollständige Systemupgrades erforderlich sind. Diese Funktionen müssen so konzipiert sein, dass sie sich schnell weiterentwickeln und Schutzmaßnahmen unabhängig von größeren Software- oder Hardware-Aktualisierungszyklen aktualisieren können.

Das richtige Gleichgewicht zwischen grundlegenden Kontrollen und anpassungsfähigen Echtzeit-Abwehrfunktionen

Stärkung der Grundlagen

Phishing-sichere MFA, Zero Trust, Least-Privilege (einschließlich für KI-Agents), diszipliniertes Patch-Management und vollständige Ressourcentransparenz.

Beseitigung struktureller Risiken

Entfernung von End-of-Life-Systemen. Ersetzung durch moderne Plattformen mit Arbeitsspeichersicherheit und Exploit-Abwehr. Sicherstellung kontinuierlicher Upgrade-Möglichkeit.

Automatisierung in Maschinengeschwindigkeit

Investition in automatisierte Erkennung, Vorselektierung und Eindämmung. Rein manuelle Reaktionsmodelle können nicht mit der Geschwindigkeit KI-gestützter Angriffe Schritt halten.

Integration aktiver Abwehrmechanismen

Platzierung von Schutzmaßnahmen im Workload, auf dem Gerät und im Datenverkehrspfad – eBPF-Laufzeitkontrollen, Inline-Durchsetzung und unabhängig aktualisierbare Exploit-Schutzschilde.

Nutzung von KI in der Abwehr

Nutzung von KI für Threat Hunting, Konformitätstests, digitale Zwillinge und Validierung zur Verkürzung von Bereitstellungszyklen von Monaten auf Tage.

Unternehmen sollten **KI-Funktionen auch für ihre eigene Abwehr nutzen**. Kontinuierliches internes Threat Hunting, unterstützt durch dieselben leistungsstarken Modelle, die Angreifer nutzen, ist eine der wichtigsten Funktionen für erfolgreiche Verteidiger. KI-gesteuerte Konformitäts- und Abnahmeprüfungen können arbeitsintensive manuelle Verifizierungen durch automatisierte Intelligence mit hoher Verarbeitungsgeschwindigkeit ersetzen. Dadurch lassen sich komplexe Testfälle generieren, die Fälle einschließen, die menschliches Testpersonal oft übersieht. In kritischen Umgebungen können KI-gestützte digitale Zwillinge Produktionsnetzwerke in großem Maßstab simulieren und überprüfen, ob Updates in Übereinstimmung mit strengen Sicherheitsprotokollen und Performance-Benchmarks durchgeführt werden, ohne die Stabilität von Live-Umgebungen zu gefährden. Durch die Integration von KI in die Abnahme- und Validierungsphasen können Engpässe bei der Bereitstellung erheblich reduziert werden, sodass der **Übergang von der Code-Fertigstellung zur Bereitstellung vor Ort von Monaten auf Tage verkürzt wird**.

In diesem neuen Umfeld erfolgreich zu sein, erfordert letztlich einen doppelten Fokus: auf die disziplinierte Durchführung grundlegender Kontrollen einerseits und auf adaptive, integrierte Sicherheitsfunktionen in Echtzeit andererseits. Unternehmen, die bestehende Risiken rigoros reduzieren, ihre Infrastruktur modernisieren, vom Vorhandensein von Sicherheitsverletzungen ausgehen und auf aktive Abwehrmodelle setzen, sind am besten aufgestellt, um die Geschwindigkeit und das Ausmaß KI-gestützter Bedrohungen zu bewältigen.

Fazit

Diese Veränderungen werden kommen. Verteidiger müssen die von ihnen verteidigte Umgebung nüchtern betrachten und damit beginnen, sie so umzugestalten, dass sie in einer von KI-gestützten Bedrohungen geprägten Welt bestehen kann. Bestehendes Wissen ist nach wie vor wichtig, muss jedoch durch moderne, innovative Abwehrmechanismen, Netzwerke mit herausragender Transparenz und angemessenen Einsatz von KI-Agents kombiniert werden, um Menschen dabei zu unterstützen, die von ihnen geschützten Umgebungen zu verteidigen.