Trend Research

TREND MICRO

SECURITY FÜR AI: EIN LEITFADEN

KI-Anwendungen richtig absichern

Künstliche Intelligenz hält zunehmend Einzug in Geschäftsprozesse. Dadurch wird sie auch zum attraktiven Ziel für Cyberkriminelle. Wie können Unternehmen Risiken minimieren und ihre Anwendungen schützen? Ein Schritt-für-Schritt-Ratgeber.

Nachdem sich der erste Hype um ChatGPT & Co. gelegt hat, ist KI mittlerweile im Unternehmensalltag angekommen. Während die Entwicklung in rasantem Tempo voranschreitet, stehen Security-Teams vor der Herausforderung, die neue Technologie richtig abzusichern. Denn KI-Anwendungen vergrößern die Angriffsfläche, bergen Schwachstellen und bringen inhärente Risiken mit sich. Ein durchdachtes Security-Konzept ist unverzichtbar, um KI-Vorteile zu nutzen und Innovationen sicher voranzutreiben. Das gilt umso dringlicher, je stärker die neue Technologie in kritische Geschäftsprozesse integriert wird.

Dabei geht es nicht nur um die technische Dimension, sondern auch um wirtschaftliche und ethische Risiken. Denn kompromittierte KI-Anwendungen könnten sensible Daten preisgeben, Geschäftsentscheidungen beeinflussen, Produktionsprozesse verändern oder sogar Menschen gefährden. Nicht umsonst schreibt der neue EU AI Act in Artikel 15 angemessene Maßnahmen zur Cybersicherheit für KI-Anwendungen mit hohem Risiko vor.

Die Bedrohungslage: Häufige Angriffe auf KI-Systeme

Cyberkriminelle greifen KI-Anwendungen in jeder Lebenszyklus-Phase an, von der Datenbeschaffung über das Training bis hin zur Inferenz. Daher ist es wichtig, Applikationen ganzheitlich abzusichern und Security von Anfang an auf allen Ebenen zu integrieren. Häufige Angriffsszenarien sind zum Beispiel Data Poisoning, Prompt Injection und Model Inversion/Extraction.

- **Data Poisoning**
  Beim sogenannten Data Poisoning schleusen Angreifer schädliche oder verfälschte Informationen in die Trainingsdaten ein. Dadurch wird das KI-Modell kompromittiert, weicht vom vorgesehenen Verhalten ab und liefert fehlerhafte oder gar gefährliche Ergebnisse. Außerdem könnten Cyberkriminelle auf diese Weise auch eine versteckte Hintertür einbauen, die später bei einem bestimmten Auslöser das Verhalten des KI-Modells verändert.

- **Prompt Injection**
  Prompt Injection ist ein beliebtes Angriffsszenario während der Inferenzphase, bei dem Cyberkriminelle das KI-Modell mithilfe von gezielten Eingaben angreifen. So versuchen sie zum Beispiel, vertrauliche Informationen zu extrahieren, integrierte Sicherheitsrichtlinien auszuhebeln oder Schadcode einzuschleusen. Besonders gefährlich: Wenn die manipulativen Prompts aus unsichtbaren Unicode-Zeichen bestehen, sind sie in der Nutzeroberfläche nicht zu sehen. Von der KI-Anwendung werden sie aber trotzdem verarbeitet.

- **Model Inversion/Model Extraction**
  Model Inversion zielt darauf ab, durch systematische Abfragen Rückschlüsse auf die ursprünglichen Trainingsdaten zu ziehen – etwa um persönliche oder vertrauliche Informationen auszulesen. Model Extraction geht noch einen Schritt weiter: Hier versuchen Angreifer, das KI-Modell zu kopieren. Dafür analysieren sie dessen Verhalten, sammeln Ein- und Ausgabepaare und trainieren ihr eigenes Modell damit.

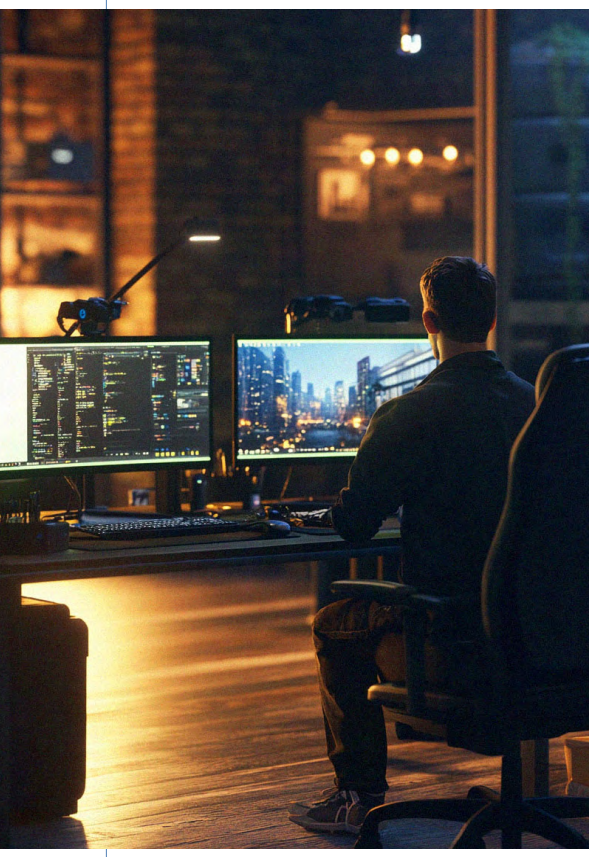
Die Angriffsfläche: Alle Ebenen der KI-Applikations-Architektur

Um KI-Systeme richtig abzusichern, müssen Unternehmen die gesamte Angriffsfläche berücksichtigen. Dafür ist es zunächst wichtig, sich bewusst zu werden, wie die Applikationsarchitektur aufgebaut ist. Eine LLM-Applikation (Large Language Model) besteht zum Beispiel aus folgenden Ebenen:

Application Services

Die Application Services bündeln alle grundlegenden Funktionen und Logiken – von der Nutzer-Authentifizierung über das Session Management und Logging bis hin zur API-Integration. Sie orchestriert die Interaktion zwischen den LLM-Agenten und anderen Komponenten, darunter Datenbanken und externe Datenquellen.

LLM-Agenten



Die LLM-Agenten fungieren quasi als Übersetzer zwischen dem LLM und anderen Komponenten. Sie bereiten eingehende Daten strukturiert auf, damit das Modell sie möglichst präzise verarbeiten kann. Außerdem managen die Agenten auch den Output, den das Modell generiert, und passen ihn an die spezifischen Anforderungen der Anwendung sowie den jeweiligen Nutzungskontext an.

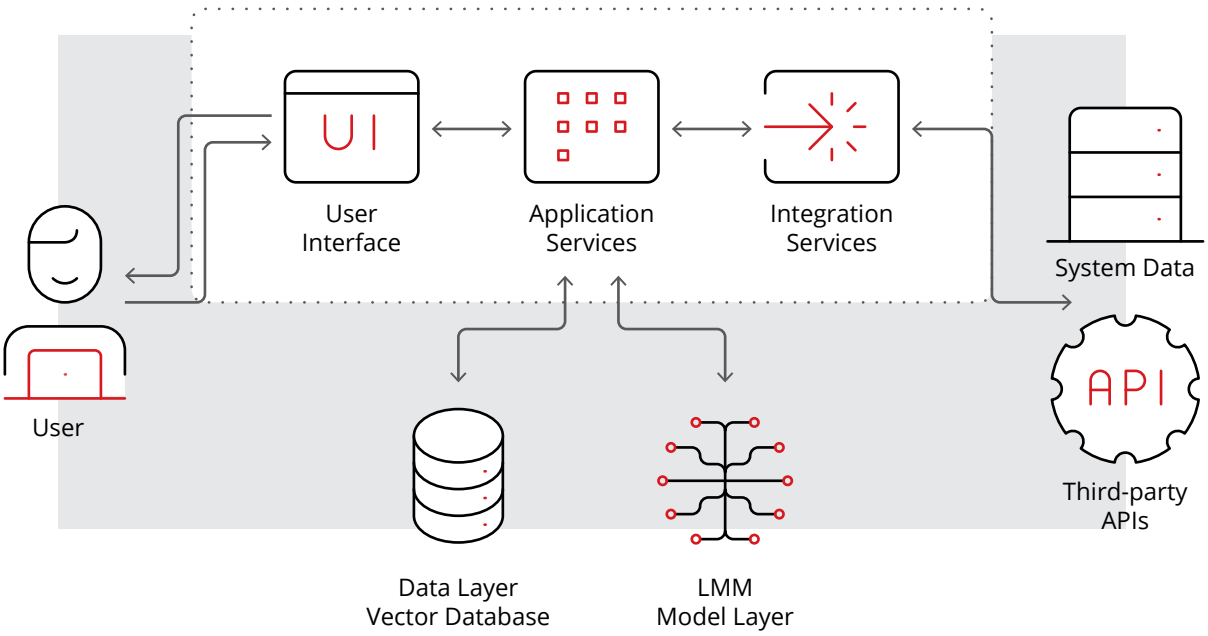
Large Language Model

Die LLM-Komponente bildet das Herzstück der Anwendung und ist maßgeblich für deren intelligente Funktionalität verantwortlich. Dabei kann es sich entweder um ein externes Modell wie GPT oder Gemini oder um eine interne, selbst gehostete Lösung handeln. Das LLM wird mit umfangreichen Daten trainiert und feinjustiert, damit es präzise und kontextgerechte Antworten liefert. Außerdem kann es mit verschiedenen Datenquellen interagieren.

ANGRIFFSVEKTOREN ENTLANG DES KI-LEBENSZYKLUS

Die Angriffsfläche eines KI-Systems umfasst sämtliche potenziellen Einstiegspunkte, über die ein Angreifer versuchen kann, das System zu kompromittieren. In der Praxis ergeben sich daraus viele verschiedene Bedrohungsszenarien entlang des gesamten Lebenszyklus einer KI-Anwendung. Hier ein Überblick über mögliche Angriffsvektoren:

1. **Daten-Inputs:** KI-Modelle sind stark datengetrieben. Indem Angreifer Eingabedaten manipulieren, können sie falsche Antworten und Fehlentscheidungen provozieren.
2. **Modelltraining:** Wenn Cyberkriminelle Zugang zu den Trainingsdaten gewinnen, können sie das KI-Modell bereits in der Trainingsphase mit Data Poisoning kompromittieren.
3. **APIs und Schnittstellen:** Cyberkriminelle können exponierte oder schlecht gesicherte APIs ausnutzen, etwa um bösartige Anfragen an das KI-Modell zu schicken oder Injection-Angriffe durchzuführen.



AI Application Framework

4. **Model-Deployment:** Mit Techniken wie Model Extraction oder Model Inversion können Angreifer versuchen, Daten zu extrahieren.

5. **Benutzerschnittstellen:** Auch Frontend-Applikationen, die KI-Modelle nutzen, sind angreifbar. Cyberkriminelle könnten sich unautorisierten Zugang verschaffen oder bösartige Skripte einschleusen.

6. **Third-Party Libraries:** Viele KI-Anwendungen nutzen externe Bibliotheken und Frameworks. Wenn Cyberkriminelle Schwachstellen in diesen Komponenten ausnutzen, ist die gesamte Applikation gefährdet.

7. **Cloud-Infrastruktur:** KI-Modelle sind häufig auf einer Cloud-Plattform gehostet. Unsichere Konfigurationen in der Cloud-Umgebung können unautorisierten Zugriff erlauben oder Denial of Service-Attacken begünstigen.

6 SCHRITTE, UM KI-APPLIKATIONEN RICHTIG ABZUSICHERN

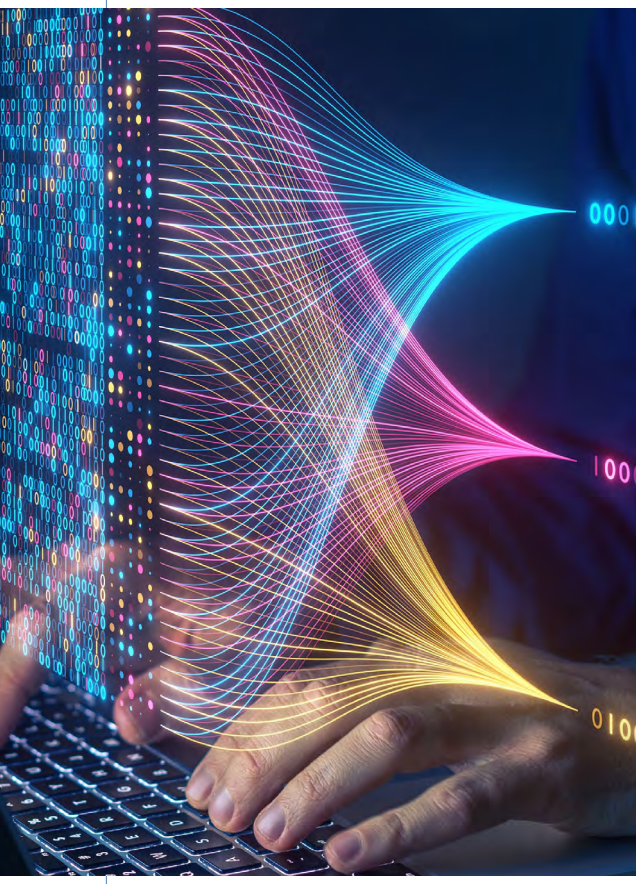
Indem Unternehmen umfassende Visibilität über die Angriffsfläche und potenzielle Bedrohungen gewinnen, können sie Schwachstellen identifizieren, Risiken managen und geeignete Sicherheitsmaßnahmen implementieren. Die folgenden sechs Schritte dienen als Blaupause für eine ganzheitliche KI-Sicherheitsstrategie:

1. Schützen Sie Ihre Daten

KI-Anwendungen basieren auf Daten. Diese angemessen zu schützen ist daher wichtig, um die Verlässlichkeit und Funktion des KI-Modells sicherzustellen. Die erste Herausforderung besteht darin, Transparenz über den Datenbestand zu gewinnen und Daten gemäß ihrer Kritikalität zu klassifizieren. Dafür eignet sich Data Security Posture Management (DSPM). Je nach Sensibilität und gesetzlichen Anforderungen können Unternehmen dann das passende Schutzniveau für jede Kategorie etablieren. Außerdem ermöglicht DSPM die Implementierung von strengen Zugriffskontrollen und das Monitoring sämtlicher Dateninteraktionen. Die Technologie überwacht sensible Daten kontinuierlich, macht potenzielle Angriffspfade sichtbar und korreliert diese Informationen. So können Unternehmen Risiken priorisieren und schnell auf Bedrohungen reagieren.

2. Schützen Sie Ihre KI-Modelle

KI-Modelle werden zunehmend in Container-Umgebungen bereitgestellt, da diese eine sehr gute Skalierbarkeit, Flexibilität und einfache Integration bieten. Solche Container sind ein beliebtes Angriffsziel und erfordern robuste Sicherheitsmaßnahmen. Dazu gehören folgende Komponenten:



- **Echtzeitmonitoring der Container Runtime.** Indem Sie das Laufzeitverhalten kontinuierlich überwachen, können Sie Anomalien schnell aufdecken.
- **Regelmäßige Schwachstellen-Scans** sowohl vor dem Deployment als auch während der Runtime, um Sicherheitslücken zu identifizieren.
- **Netzwerksegmentierung,** um KI-Container von anderen Bereichen der Infrastruktur zu separieren. So können Sie das Schadensmaß bei einem potenziellen Angriff eingrenzen.
- **Ein umfassendes, detailliertes Logging aller Container-Aktivitäten,** etwa durch XDR (Extended Detection and Response). Diese Daten ermöglichen nicht nur eine schnelle Reaktion, sondern erleichtern im Falle eines Cyberangriffs auch die forensische Analyse.

3. Sichern Sie Ihre KI-Infrastruktur

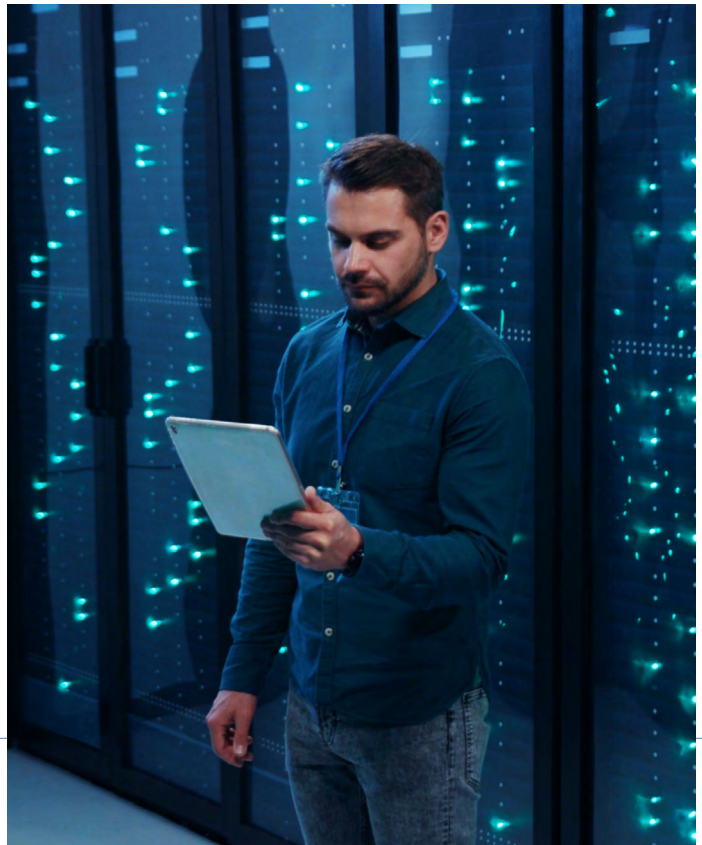
Um die Sicherheit, Compliance und Integrität der KI-Infrastruktur zu gewährleisten, empfiehlt sich AI Security Posture Management (AI-SPM). Im ersten Schritt geht es darum, umfassende Visibilität in der KI-Infrastruktur zu schaffen, einschließlich der Komponenten, Konfigurationen und Datenflüsse in der Cloud-Umgebung. Kontinuierliches Monitoring ermöglicht ein proaktives Risikomanagement, während regelmäßige Schwachstellen-Scans Sicherheitslücken aufdecken. Außerdem spielt AI-SPM mit CSPM (Cloud Security Posture Management) und KSPM (Kubernetes Security Posture Management) zusammen – etwa um unsichere Konfigurationen in Cloud-Umgebungen zu identifizieren oder die Compliance mit Regularien wie der DSGVO, HIPAA oder PCI-DSS sicherzustellen.

4. Schützen Sie Ihre Nutzer und lokalen KI-Anwendungen

Sensibilisieren Sie Ihre Mitarbeiter für KI-Risiken und einen sicheren Umgang mit der Technologie. Dazu gehören zum Beispiel Awareness-Trainings und eine KI-Policy, die Dos und Don’ts beim KI-Einsatz im Unternehmen klar definiert. Um Datenrisiken zu mindern und sensible Informationen besser zu schützen, setzen viele Unternehmen auf lokale KI-Installationen. Da auch diese anfällig für Cyberangriffe wie Data Poisoning sein können, sollten sie kontinuierlich überwacht werden.

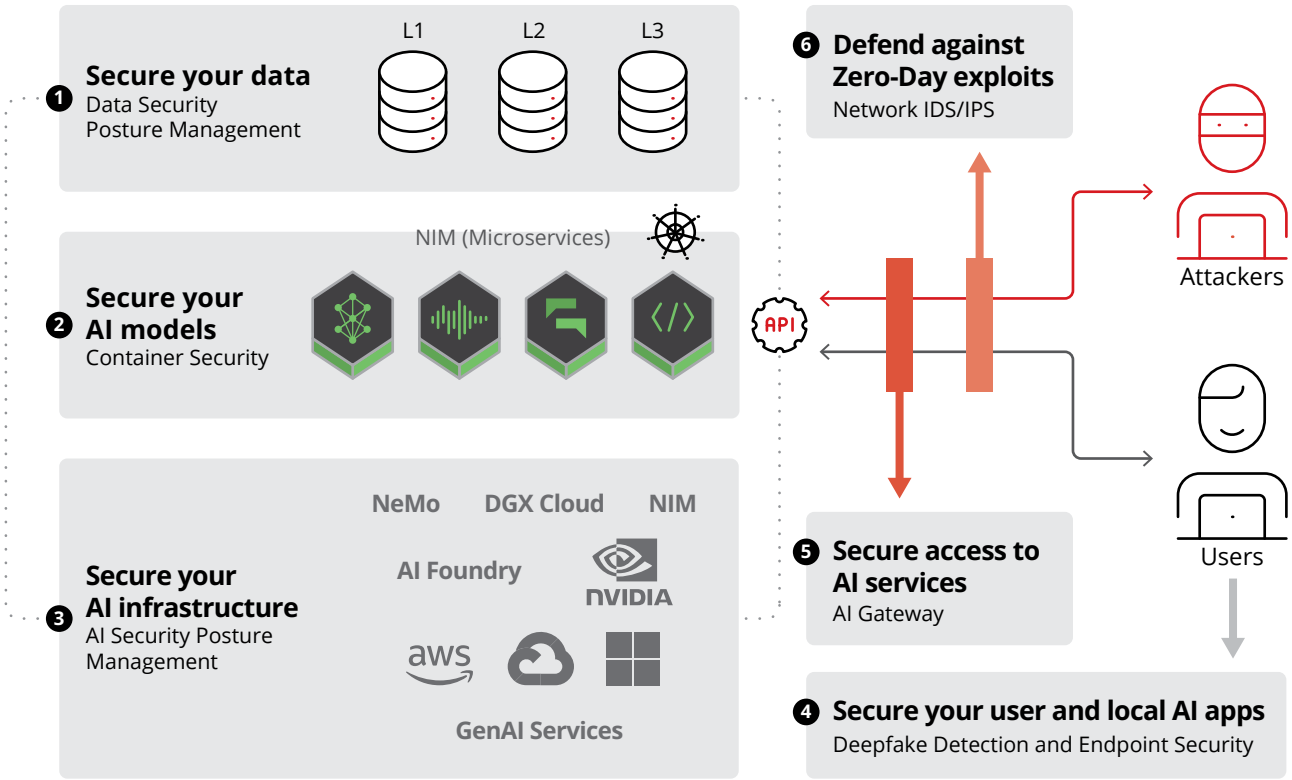
5. Sichern Sie den Zugriff auf KI-Anwendungen

Mit einem Zero-Trust-Ansatz sorgen Sie dafür, dass nur autorisierte Nutzer auf KI-Services zugreifen können. Dies ist insbesondere im Hinblick auf öffentliche KI-Services wichtig. Am besten lässt sich das Konzept mit einer ZTSA-Lösung umsetzen (Zero Trust Secure Access), die risikobasierte Zugriffskontrollen ermöglicht. Als Reverse-Proxy zieht sie außerdem eine zusätzliche Sicherheitsebene zwischen Anwender und KI-Services ein. Auf diese Weise lassen sich sowohl Input als auch Output der KI-Anwendungen kontrollieren, etwa um bösartige Eingaben oder die Preisgabe sensibler Daten zu verhindern. Darüber hinaus macht ZTSA die KI-Nutzung in der gesamten Organisation transparent und deckt Schatten-KI auf, die von Mitarbeitern in Eigenregie genutzt wird.



6. Verteidigen Sie sich gegen Zero Day Exploits

Zero Day Exploits zielen darauf ab, bisher unbekannte Sicherheitslücken auszunutzen. Da für solche Schwachstellen noch keine Patches existieren, gelten sie als besonders gefährlich. Um KI-Anwendungen vor Zero Day Exploits zu schützen, sind Security-Maßnahmen auf Netzwerkebene wie Intrusion Detection- oder Intrusion Prevention-Systeme (IDS/IPS) wichtig. Indem diese den Datenverkehr in Echtzeit überwachen, können sie verdächtige Aktivitäten und potenzielle Zero-Day-Exploits schnell erkennen und automatisiert reagieren. Dafür kommen globale Threat Intelligence und KI-gestützte Analyse-Techniken zum Einsatz, die lernen, zwischen normalem und verdächtigem Verhalten zu unterscheiden. Eine wichtige Funktion ist zudem Virtual Patching: Diese Technologie deckt ungepatchte Schwachstellen auf Netzwerkebene ab, sodass sie nicht mehr von außen angreifbar sind.



Security for AI Blueprint

Fazit

KI-Sicherheit ist kein einmaliges Projekt, sondern ein kontinuierlicher Prozess. Unternehmen müssen sowohl technische als auch organisatorische Maßnahmen integrieren und regelmäßig auf den Prüfstand stellen. Denn während sich die neue Technologie rasant weiterentwickelt, verändern sich auch die Sicherheitsanforderungen. Entscheidend ist, die gesamte Angriffsfläche entlang des KI-Lebenszyklus zu betrachten und Security von Anfang an in neue Projekte zu integrieren. Dafür müssen Sicherheits- und KI-Teams eng zusammenarbeiten und eine DevSecOps-Kultur etablieren.

Möchten Sie weitere Einblicke wie diesen?

www.trendmicro.com/de_de/business/ai-cybersecurity.html

SECURITY FÜR AI: EIN LEITFADEN

TREND MICRO™ | Global Leader in Cybersecurity

Trend Micro, a global cybersecurity leader, helps make the world safe for exchanging digital information. Fueled by decades of security expertise, global threat research, and continuous innovation, Trend Micro's AI-powered cybersecurity platform protects hundreds of thousands of organizations and millions of individuals across clouds, networks, devices, and endpoints. As a leader in cloud and enterprise cybersecurity, Trend's platform delivers a powerful range of advanced threat defense techniques optimized for environments like AWS, Microsoft, and Google, and central visibility for better, faster detection and response. With 7,000 employees across 70 countries, Trend Micro enables organizations to simplify and secure their connected world.

For more information visit www.TrendMicro.com.