eBook

THALES
CYBERSECURITY

KI von innen heraus sichern: Ein Leitfaden zu den 7 größten LLM-Risiken

cpl.thalesgroup.com

Inhalt

Warum KI das Bedrohungsmodell verändert	4
Die 7 Risiken, die jedes KI-Team verstehen muss	5
Der gemeinsame Nenner: Daten sind immer das Ziel	8
Ein konzeptioneller Ansatz für KI-Sicherheit	9
Schutz vor zukünftigen Datenrisiken	10
Über Thales	11

Einführung

Jedes KI-System läuft auf der Grundlage von Daten und häufig auch auf der Grundlage der sensibelsten Daten, die eine Organisation besitzt. Trainingsdatensätze, Abruf-Pipelines und Agenten-Frameworks erfordern einen umfassenden Zugriff auf interne Wissensquellen.

Laut dem Thales Data Threat Report 2026 berichten 61 % der Organisationen, dass ihre KI-Anwendungen bereits aktiv ins Visier genommen werden, wobei sensible Daten das primäre Ziel sind. Die Angriffsfläche ist neu, nicht aber das Ziel.

Der Wert der KI hängt vollständig von Daten ab, und diese Daten sind nun auf einer Angriffsfläche sichtbar, die vor 3 Jahren noch nicht existierte. Diese Angriffsoberfläche ist komplex und umfasst Eingabeaufforderungen, Vektordatenbanken, Agentenframeworks, Abrufpipelines und Modell-APIs, die nun neue Wege zu sensiblen Daten eröffnen.

Dieser Leitfaden richtet sich an Führungskräfte, die verstehen müssen, was auf dem Spiel steht. Er beschreibt sieben kritische LLM-Risiken, erläutert, wie sich diese in konkrete organisatorische Risiken niederschlagen, und bietet einen Rahmen für fundierte Entscheidungen zur KI-Sicherheit.

Das macht KI-Infrastruktur zu einem so überzeugenden Ziel: kompromittiere das Modell, und du hast eine potenzielle Schnittstelle zu allem, auf das die KI Zugriff hat.

Warum KI das Bedrohungsmodell verändert

Obwohl KI die Form einer Anwendung annehmen kann, unterscheiden sich KI-Risiken von der traditionellen Anwendungssicherheit. Traditionelle Sicherheit setzt klare Grenzen voraus: definierte Eingaben, deterministische Anwendungslogik und defensive Kontrollen an bekannten Kontrollpunkten.

Große Sprachmodelle (LLMs) arbeiten jedoch in durchlässigen und dynamischen Umgebungen. Eingabeaufforderungen, externe Datenquellen, APIs und autonome Agenten interagieren kontinuierlich mit dem Modell.

Das Ergebnis ist eine Anwendungsoberfläche, die sich eher anhand des Kontexts und des Datenflusses als anhand statischer Logik verändert.

Sicherheitsteams müssen sich auch damit auseinandersetzen, wie schnell sich die KI-Technologie verändert. Das KI-Ökosystem entwickelt sich schnell weiter, wobei jedes Quartal neue Frameworks, Tools und Architekturen entstehen. Tatsächlich geben laut dem Data Threat Report 70 % der Unternehmen die rapide Entwicklung als ihr größtes KI-Sicherheitsbedenken an.

Dies führt zu mehreren entscheidenden Unterschieden gegenüber herkömmlichen Softwaresystemen:

Dynamische Eingaben:
Prompts und kontextbezogene Daten beeinflussen das Verhalten des Systems in Echtzeit.

Nicht-deterministische Ergebnisse:
Die Modellreaktionen variieren, was die Validierung und Kontrolle erschwert.

Expandierende Ökosysteme:
Durch Retrieval-Augmented Generation (RAG), Plugins und Agenten-Frameworks werden neue Abhängigkeiten und Integrationen eingeführt

Die rapiden KI-Veränderungen machen statische Kontrollen weniger effektiv und erhöhen das Risiko fragmentierter Sicherheitsentscheidungen. Dies ist der Grund, warum die OWASP Top 10 für LLMs so wichtig sind. Sie ordnet die wichtigsten Risiken ein, die in der KI von Unternehmen auftreten – von Promt Injektion bis hin zu Datenlecks – und hilft Teams dabei, von abstrakten Befürchtungen zur praktischen Bewertung überzugehen.

7 Risiken, die jedes KI-Team verstehen muss

Nicht jedes KI-Risiko lässt sich gleichermaßen angehen. Die Top 10 von OWASP für LLMs ist eine umfassende Taxonomie, aber mehrere Risiken, darunter Modellhalluzinationen, Schwachstellen in der Lieferkette und Verhalten der Agents, hängen weitgehend von Faktoren ab, die außerhalb der direkten Kontrolle einer Organisation liegen. Dieser Leitfaden konzentriert sich auf die sieben Risiken, die innerhalb des Enterprise-Security-Perimeters liegen: die Daten-, Anwendungs- und Identitätsebenen, auf denen Kontrollen entworfen, durchgesetzt und gemessen werden können.

Dies sind die Risiken, bei denen Sicherheitsteams jetzt handeln können, gruppiert in 3 Kategorien: Integrität von Ein- und Ausgabedaten, Sicherheitslücken auf Datenebene und Zugriff auf die Infrastruktur.

Input/Output-Integrität

01

Prompt-Injektion

nutzt das grundlegendste Merkmal von LLMs aus: Sie verhalten sich entsprechend der Anweisungen und des Kontextes, den sie erhalten. Angreifer betten versteckte Befehle in Benutzeraufforderungen oder abgerufenen Inhalt ein, um das beabsichtigte Verhalten des Modells zu überschreiben.

Im Erfolgsfall könnte das Modell Schutzmaßnahmen ignorieren, sensible Daten offenlegen oder unerwünschte Aktionen in verbundenen Systemen auslösen.

Frage:

Wenn Ihr KI-System eine Entscheidung getroffen hätte, die einen Kunden oder Partner beeinträchtigt hat, könnten Sie erklären, wie es dazu gekommen ist und nachweisen, dass niemand in diesen Prozess eingegriffen hat?

02

Unsachgemäße Verarbeitung der Ausgabedaten

tritt auf, wenn KI-Antworten ohne Überprüfung als vertrauenswürdige Ausgabe behandelt werden. Werden manipulierte Ausgaben direkt in Skripte, APIs oder Datenbankabfragen eingespeist, können sie unbeabsichtigte Kommandos auslösen oder nachgelagerte Systeme gefährden.

03

System-Prompt-Leck

tritt auf, wenn Angreifer die internen Anweisungen extrahieren, die das Verhalten eines Modells steuern. Diese Prompts enthalten häufig Regeln, Sicherheitsvorkehrungen oder Konfigurationsdetails. Werden sie offengelegt, geben sie Aufschluss darüber, wie das System funktioniert, und helfen Angreifern dabei, effektivere Prompt-Injektionen zu entwickeln oder Sicherheitsvorkehrungen zu umgehen

Risiken auf der Datenebene

04

Offenlegung sensibler Informationen

ist die unmittelbarste Folge der Abhängigkeit von KI von Unternehmensdaten. Modelle, die mit internen Wissensquellen trainiert oder erweitert wurden, können unbeabsichtigt personenbezogene Daten, Anmeldedaten oder geschützte Informationen in ihren Ausgaben offenlegen. Das Risiko steigt noch, wenn man bedenkt, dass laut dem „2026 Data Threat Report" nur 47 % der sensiblen Cloud-Daten verschlüsselt sind.

05

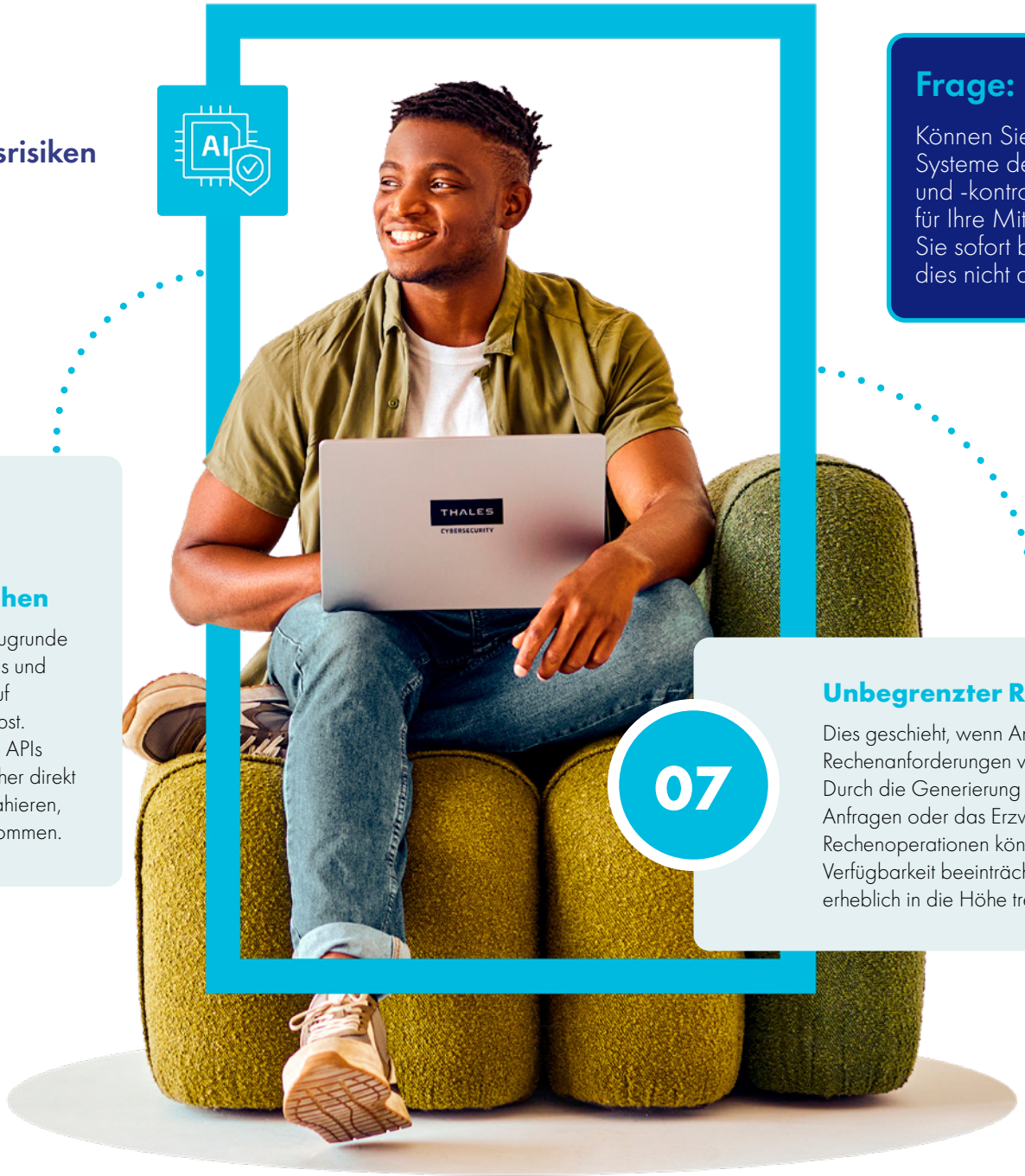
Daten- und Modellvergiftung

tritt ein, wenn Angreifer manipulierte Inhalte in Trainingsdatensätze oder Wissensdatenbanken einschleusen. Dies kann das Modellverhalten verändern, Verzerrungen verursachen oder versteckte Auslöser einbetten, die unter bestimmten Bedingungen aktiviert werden, wodurch KI-Modelle im Grunde zu einer Insider-Bedrohung werden. Der Unterschied zu herkömmlichen Insider-Bedrohungen liegt im Ausmaß: Ein böswilliger Insider beeinflusst nur das, was er erreichen kann; ein kompromittiertes Modell beeinflusst alles, was es erreichen kann.

Frage:

Wenn ein Prüfer wissen möchte, auf welche Daten Ihre KI-Systeme in den letzten 30 Tagen zugegriffen haben und wann dies geschah, könnte Ihr Team diesen Nachweis dann zweifelsfrei erbringen?

Infrastruktur- und Zugriffsrisiken



Frage:

Können Sie nachweisen, dass Ihre KI-Systeme den gleichen Zugriffsrichtlinien und -kontrollen unterliegen, die auch für Ihre Mitarbeiter gelten, und dass Sie sofort benachrichtigt werden, falls dies nicht der Fall ist?

06

Vektor- und Einbettungsschwächen

beruhen auf der Abhängigkeit der KI von der zugrunde liegenden Infrastruktur (Vektordatenbanken, APIs und Rechenumgebungen), die oft weniger streng auf ihre Sicherheit geprüft wird als die Modelle selbst. Schwache Zugriffskontrollen oder ungeschützte APIs können es Angreifern ermöglichen, Vektorspeicher direkt abzufragen und sensible Informationen zu extrahieren, ohne jemals mit dem Modell in Berührung zu kommen.

07

Unbegrenzter Ressourcen-Verbrauch

Dies geschieht, wenn Angreifer die hohen Rechenanforderungen von KI-Modellen ausnutzen. Durch die Generierung großer Mengen an Anfragen oder das Erzwingen ressourcenintensiver Rechenoperationen können Angreifer die Verfügbarkeit beeinträchtigen oder die Betriebskosten erheblich in die Höhe treiben.

Obwohl die Risiken technisch unterschiedlich sind, verfolgen sie alle dasselbe Ziel: Angreifern den Zugriff auf sensible Daten zu ermöglichen, diese zu manipulieren oder zu stehlen.

In diesem Sinne verhalten sich KI-Systeme wie privilegierte Insider. Sie arbeiten mit einem breiten, legitimen Zugriff auf sensible Datenquellen. Sie handeln innerhalb vertrauenswürdiger Systemgrenzen. Ihre Ergebnisse wirken normal, bis etwas schief läuft. Genau darauf bezieht sich der Thales Data Threat Report 2026, wenn er KI als die neue Insider-Bedrohung einstuft: nicht, dass KI von Natur aus bösartig ist, sondern dass ein kompromittiertes, manipuliertes oder schlecht verwaltetes KI-System dasselbe strukturelle Profil aufweist wie ein vertrauenswürdiger Mitarbeiter, der die Seite gewechselt hat.

Die Angriffsfläche ist nicht nur das Modell. Es ist alles, was das Modell erreichen kann.

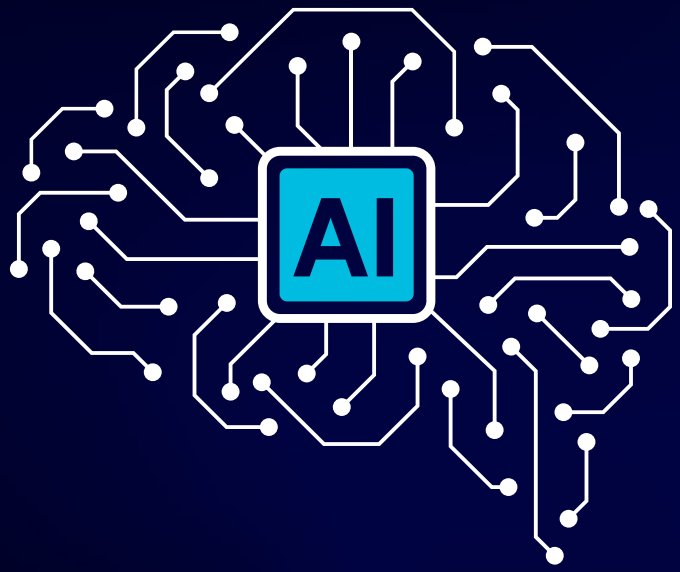
Dies spiegelt die Struktur von KI-Systemen wider; Modelle funktionieren nur, wenn sie auf große Mengen an Unternehmensdaten zugreifen können, häufig über mehrere Speicherorte und Cloud-Umgebungen hinweg. Um effektiv zu arbeiten, müssen sie tief in interne Wissensquellen integriert sein.

Diese Integration schafft Möglichkeiten für Angreifer. Einfach ausgedrückt: Wenn ein Angreifer das KI-Modell kompromittieren kann, kann er es als Brücke zu den Daten nutzen, auf die diese Modelle zugreifen.

Das Problem wird dadurch verschärft, dass viele Unternehmen wenig Einblick in ihre Daten haben. Laut dem Data Threat Report 2026 wissen nur 34 % der Unternehmen, wo sich all ihre Daten befinden, und nur 39 % können sie vollständig klassifizieren.

Ohne genaue Datenermittlung und -klassifizierung ist es unmöglich, zuverlässig zu erkennen, welche Interaktionen sensible Informationen betreffen und welche daher einen stärkeren Schutz benötigen.

Daten sind in KI-Umgebungen alles. Sie treiben KI-Modelle an und sind am meisten gefährdet. Daher müssen Security-Programme mit einem klaren Verständnis der Daten selbst beginnen. Das bedeutet, zu verstehen, wo sie sich befinden, wie sie sich bewegen und wer darauf zugreifen kann.



Ein konzeptioneller Ansatz für KI-Sicherheit

Die Herausforderung für Sicherheitsverantwortliche liegt nicht in einem Mangel an KI-Sicherheitstools, sondern in einem Mangel an Kohärenz. Laut dem „2026 Data Threat Report" setzen Unternehmen durchschnittlich sieben Datenschutz-Tools ein, doch nur 39 % sind davon überzeugt, dass diese wirksam sind. Mehr Tools ohne ein einheitliches Rahmenkonzept verringern das Risiko nicht. Sie verteilen es lediglich neu.

Aus diesem Grund benötigen Sie einen einheitlichen Ansatz, der die KI-Sicherheit über drei Kernbereiche hinweg organisiert.

Daten-Sicherheit

Der erste Bereich konzentriert sich auf den Schutz der Daten, auf die KI-Systeme angewiesen sind. Dazu gehört, sensible Daten zu identifizieren, sie zu verschlüsseln und den Zugriff darauf zu kontrollieren.

In KI-Umgebungen helfen diese Maßnahmen, Risiken zu begrenzen, wenn Angreifer Zugriff auf Abruf-Pipelines, Trainingsdaten oder Vektorspeicher erlangen. Eine starke Datensicherheit reduziert die Auswirkungen von Risiken wie der Offenlegung sensibler Informationen, Datenvergiftung und Datenexfiltration.

Applikations-Sicherheit

Der zweite Bereich konzentriert sich darauf, wie die KI-Anwendung Eingaben und Ausgaben verarbeitet. Modelle rufen Daten ab und reagieren auf Eingabeaufforderungen und den Kontext, weshalb die Validierung von Ein- und Ausgaben von entscheidender Bedeutung ist. Kontrollen auf Anwendungsebene tragen dazu bei, vor dem Einschleusen von Eingabeaufforderungen zu schützen, die Offenlegung von System-Eingabeaufforderungen zu verhindern und sicherzustellen, dass Modellantworten keine unsicheren nachgelagerten Aktionen auslösen können.

Identität und Zugriffssicherheit

Der letzte Bereich konzentriert sich darauf, wer – und was – mit KI-Systemen interagieren darf. Dazu gehören Nutzer, Dienste, Agenten und Anwendungen, die auf Modelle oder die ihnen zugrunde liegenden Daten zugreifen. Strenge Identitäts- und Zugriffskontrollen stellen sicher, dass Berechtigungen begrenzt und überwacht werden. Dies hilft, unbefugte Interaktionen – sei es durch Menschen oder Maschinen – mit KI-Infrastruktur, Abrufsystemen und sensiblen Datenquellen zu verhindern.

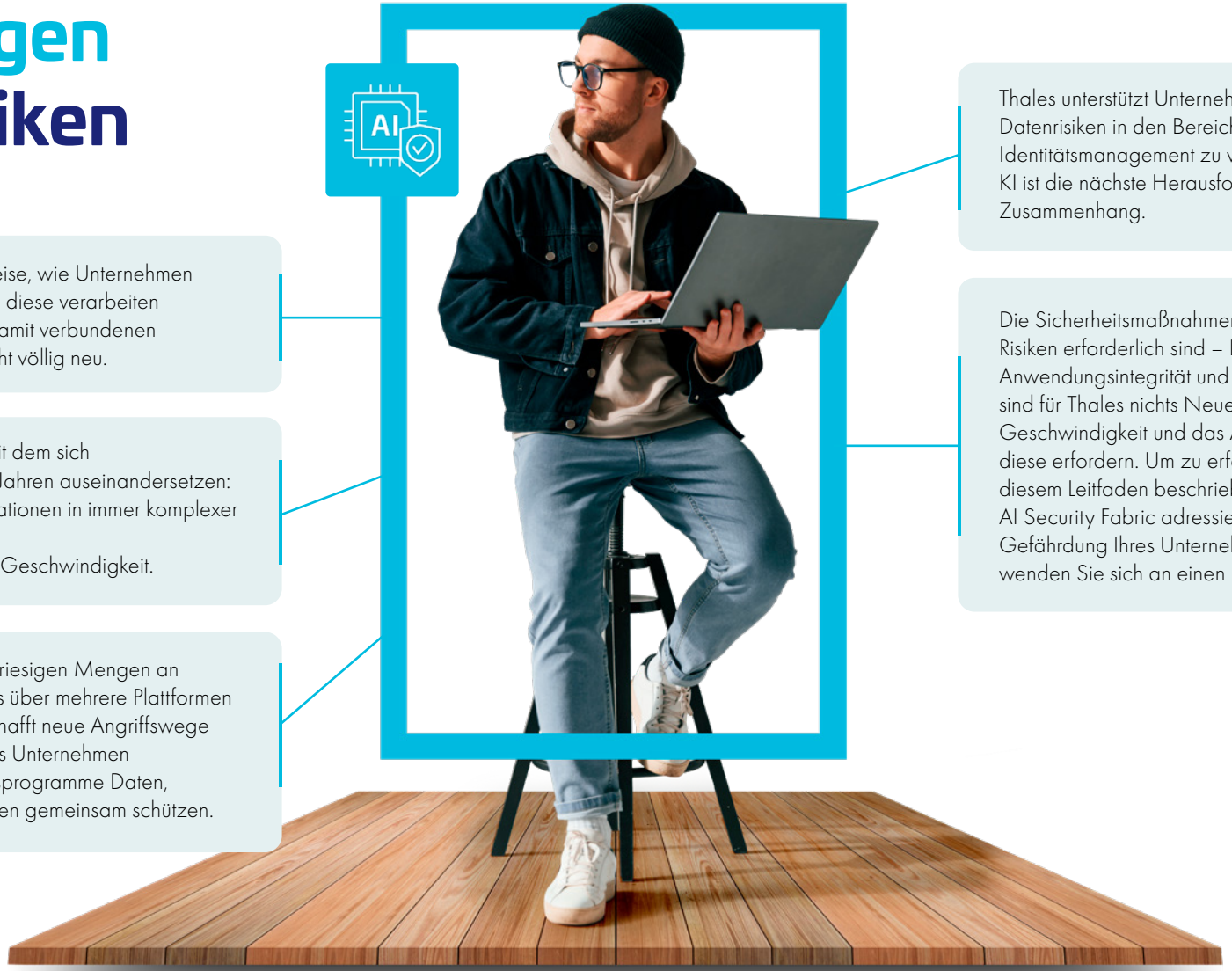
Zusammen bieten diese drei Bereiche einen praktischen Ansatz zum Verständnis und Management von KI-Risiken

Schutz vor zukünftigen Datenrisiken

KI verändert die Art und Weise, wie Unternehmen auf Informationen zugreifen, diese verarbeiten und generieren. Doch die damit verbundenen Herausforderungen sind nicht völlig neu.

KI verschärft ein Problem, mit dem sich Sicherheitsteams schon seit Jahren auseinandersetzen: den Schutz sensibler Informationen in immer komplexer werdenden Umgebungen.
Der Unterschied liegt in der Geschwindigkeit.

KI-Systeme interagieren mit riesigen Mengen an Unternehmensdaten, oftmals über mehrere Plattformen und Dienste hinweg. Das schafft neue Angriffswege und erfordert vor allem, dass Unternehmen überdenken, wie Sicherheitsprogramme Daten, Anwendungen und Identitäten gemeinsam schützen.



Thales unterstützt Unternehmen seit Jahren dabei, Datenrisiken in den Bereichen Cloud, Compliance und Identitätsmanagement zu verstehen und zu bewältigen. KI ist die nächste Herausforderung in diesem Zusammenhang.

Die Sicherheitsmaßnahmen, die zur Abwehr dieser Risiken erforderlich sind – Datentransparenz, Anwendungsintegrität und Identitätskontrolle – sind für Thales nichts Neues. Neu ist jedoch die Geschwindigkeit und das Ausmaß, in dem KI-Systeme diese erfordern. Um zu erfahren, wie Thales die in diesem Leitfaden beschriebenen Risiken mithilfe seiner AI Security Fabric adressiert, oder um die aktuelle Gefährdung Ihres Unternehmens zu bewerten, wenden Sie sich an einen unserer Thales-Experten

Über Thales

Thales ist ein weltweit führender Anbieter im Bereich Cybersicherheit und unterstützt Unternehmen, Regierungen und die renommiertesten Organisationen der Welt dabei, kritische Anwendungen, sensible Daten, Identitäten und Software überall und in großem Maßstab zu schützen – bei höchster Rentabilität. Mit mehr als 30.000 Kunden, darunter 58 % der Fortune Global 500-Unternehmen, werden unsere Lösungen in 148 Ländern weltweit eingesetzt. Durch unsere innovativen Dienstleistungen und integrierten Plattformen hilft Thales seinen Kunden dabei, Risiken besser zu erkennen, sich gegen Cyberbedrohungen zu schützen, Compliance-Lücken zu schließen und täglich für Milliarden von Verbrauchern vertrauenswürdige digitale Erlebnisse zu schaffen.

THALES

CYBERSECURITY

Kontakt

Kontaktinformationen finden Sie unter cpl.thalesgroup.com/contact-us

cpl.thalesgroup.com