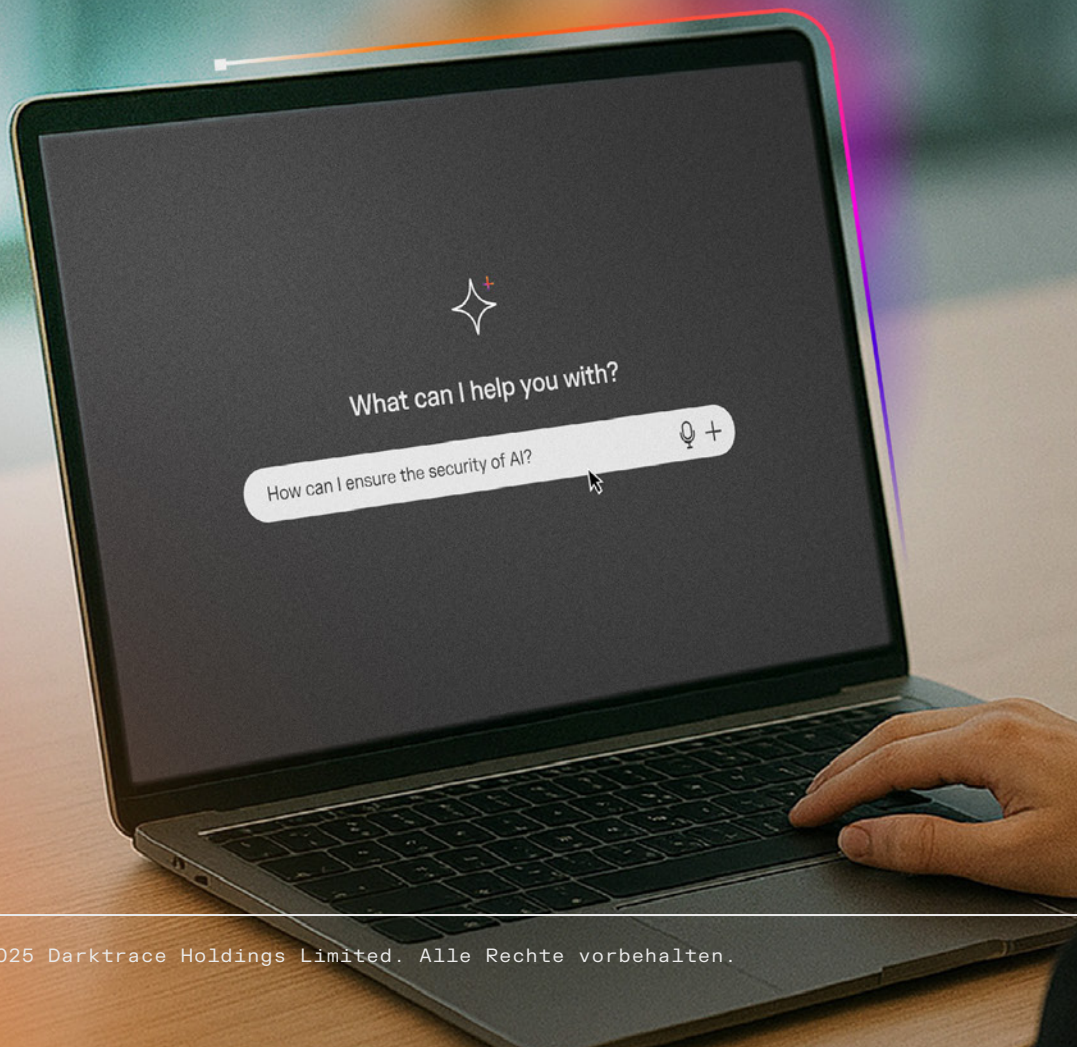


So sichern Sie KI im Unternehmen

Ein Leitfaden zur Sicherung von
Modellen, Daten und Agenten



Inhalt

02	Vorwort	07	Überprüfung von KI-generiertem Code
02	Über Darktrace	08	Sicherung der KI-Lieferkette
03	KI absichern: Verstehen, was auf dem Spiel steht	08	Einsatz von Shadow AI
04	5 Kategorien von KI-Risiken und wie man ihnen begegnet	08	KI-Anbieter, die Agenten direkt bereitstellen
05	Schutz vor Missbrauch und unerwarteten Verhaltensweisen	08	KI im Lieferantenmanagement
05	Analyse bösartiger Eingabeaufforderungen	08	Code- und Abhängigkeitsprüfung für KI-Agenten
05	Unbeabsichtigte Anwendungsfälle	09	Angebot an KI-Modellen
05	Zugriff auf den Prompt-Verlauf	09	KI-Trainingsdaten
05	Böswillige und halluzinatorische Ausgabeeffekte	10	Stärkung der Einsatzbereitschaft und Aufsicht
06	Überwachung und Steuerung von KI im Betrieb	10	Berichterstattung
06	Aktionen und Verbindungen von KI-Agenten	10	KI-Penetrationstests, Red- und Purple-Teaming
06	KI-Gateway	10	Schulungen und Übungen für Sicherheitsteams
06	KI-Agenten-Zustandserfassung	11	Ein neues Verständnis von KI-Sicherheit für die Unternehmensleitung
06	KI-Datenübertragung	12	Fazit
07	Schutz der KI-Entwicklung und -Infrastruktur		
07	KI-Sicherheitskonfiguration		
07	KI-Agenten-Architekturen		
07	laC-Scanning		

Vorwort

Wir haben von unseren Kunden erfahren und aus erster Hand beobachtet, dass es Sicherheitsverantwortlichen an einer klaren, prägnanten Definition dessen mangelt, was unter „Sicherheit für KI“ zu verstehen ist. Zwar gibt es wertvolle Ressourcen zu diesem Thema, doch haben wir bislang noch keine gefunden, die Risiken und Sicherheitsfunktionen schlüssig miteinander verknüpft. Umfassendere Rahmenwerke wie ISO/IEC 42001 liefern zwar wichtige Governance-Grundsätze, bieten jedoch keine praktischen Leitlinien, die für die operative Umsetzung von KI-Sicherheit in einem Unternehmen erforderlich sind.

Dieser Beitrag soll diese Lücke schließen. Er bietet einen herstellerneutralen Leitfaden zum Verständnis des Umfangs der Cybersicherheit im Zusammenhang mit KI und beschreibt die wichtigsten Kategorien, Risiken und Sicherheitsfunktionen, die Unternehmen bei der Einführung und Integration von KI-Technologien in ihrem gesamten Geschäftsbetrieb berücksichtigen sollten.

Über Darktrace

Darktrace ist seit über einem Jahrzehnt führend bei der Anwendung künstlicher Intelligenz im Bereich der Cybersicherheit und hat mit bahnbrechenden Innovationen die Branche grundlegend verändert. Im Jahr 2013 haben wir „Netzwerk-Erkennung und -Reaktion (NDR)“ (NDR) eingeführt, das vollständig auf KI-gestützten Erkennungsmethoden basiert und damit einen neuen Standard für die autonome Erkennung von Bedrohungen gesetzt hat.

Bis 2019 haben wir einen autonomen KI-Ermittlungsagenten eingeführt, der menschliche Ermittlungs- und Analyseverfahren mit maschineller Geschwindigkeit und in maschinellem Umfang nachbildet. Durch den frühzeitigen Einsatz fortschrittlicher Sprachmodelle, darunter Transformer-basierte Ansätze im Jahr 2021 und domänenspezifische Sprachmodelle (DSLMS) im Jahr 2023, konnten wir unsere Fähigkeit zur Analyse komplexer Cyberbedrohungen weiter stärken.

Heute ist die Darktrace-KI in rund 10.000 Unternehmen weltweit im Dauereinsatz und sorgt für bewährte Widerstandsfähigkeit in großem Maßstab. Mitte 2025 gehörten wir zu den ersten Cybersicherheitsunternehmen, die nach ISO 42001 zertifiziert wurden, was unser Engagement für eine verantwortungsvolle KI-Governance unterstreicht. Dies spiegelt nicht nur unsere Vorreiterrolle bei Innovationen im Bereich der KI wider, sondern auch unser unerschütterliches Engagement für die Sicherung der KI selbst, damit Vertrauen, Transparenz und Sicherheit stets im Mittelpunkt jedes Fortschritts stehen.



Über 10 Jahre Erfahrung in der Anwendung von KI auf Anwendungsfälle im Bereich Cybersicherheit



Eines der ersten Cybersicherheitsunternehmen, das nach ISO 42001 zertifiziert wurde



Gewinner des AI Excellence Awards für KI im Bereich Cybersicherheit (2025)



Rund 10.000 Kunden, Tendenz steigend

Sicherheit im Bereich der KI

Verstehen, was auf dem Spiel steht

Künstliche Intelligenz hält in Unternehmen in einem Tempo Einzug, mit dem nur wenige Governance- und Managementsysteme mithalten können.

Herkömmliche Sicherheitskonzepte haben Mühe, Schritt zu halten, da KI neue Komplexitäten mit sich bringt und die Angriffsfläche vergrößert. Jedes Modell, jeder Datensatz und jeder Agent sowie das komplexe Geflecht aus APIs, Integrationen von Drittanbietern und menschlichen Interaktionen, das diese Systeme antreibt, birgt neue Risiken und eröffnet neue Angriffswege.

KI verspricht für manche Produktivitäts-, Geschwindigkeits- und Effizienzsteigerungen, bringt jedoch auch eine Reihe von Unsicherheiten mit sich und erfordert insbesondere im Bereich der generativen KI und der auf GenAI basierenden Agenten ein zunehmendes Maß an Aufsicht. Generative KI, die unzuverlässige Ergebnisse oder Halluzinationen erzeugen kann, bei der die Entscheidungsfindung intransparent ist und die autonom auf eine Weise arbeitet, für deren Überwachung herkömmliche Sicherheitstools nie konzipiert wurden, erweitert das digitale Ökosystem, in dem Eigentumsverhältnisse, Kontrolle und Verantwortlichkeit oft verschwimmen.

Für Sicherheitsverantwortliche besteht die erste Herausforderung beim Schutz eines Unternehmens in Zeiten des Wandels darin, zu ermitteln, was abgesichert werden muss. KI hält über verschiedene Kanäle Einzug in Unternehmen. Die Mitarbeiter nutzen mittlerweile KI-gestützte SaaS- und Produktivitäts-Tools. Cloud- und On-Prem-Anbieter erweitern ihre Infrastruktur und ihre Dienste um integrierte KI-Funktionen. Drittanbieter und erworbene Software führen häufig KI-Funktionen ein, wodurch sich das Risiko ohne klare Kontrolle erhöht. Intern können Entwickler eigene Modelle und Agenten erstellen oder anpassen, die KI direkt in das Unternehmen integrieren.

Über alle diese Zugangspunkte hinweg erfordert der Schutz eines KI-Ökosystems Transparenz auf allen Ebenen: von der Entwicklung und dem Training bis hin zur Bereitstellung und zum Echtzeitbetrieb. Jede Phase birgt ganz eigene Risiken, von Datenvergiftung und Abhängigkeitsfehlern bis hin zu böswilligem Verhalten von Agenten oder der unbeabsichtigten Nutzung von Modellausgaben.



5 Kategorien von KI-Risiken und wie man ihnen begegnet

Jede der fünf in diesem Rahmenwerk beschriebenen Kategorien stellt eine eigene Ebene der KI-Sicherheit dar, von externen Anbietern und Entwicklungspipelines bis hin zum Live-Betrieb und der organisatorischen Steuerung.

In jedem dieser Bereiche bestimmen spezifische Funktionen, wo Risiken entstehen können und wo Sicherheitsmaßnahmen eingesetzt werden können, um das Risiko zu verringern. Im folgenden Abschnitt werden diese Funktionen im Detail erläutert, wobei jeweils die damit verbundenen Risiken sowie die Sicherheitsfunktionen, die zu deren Bewältigung dienen, aufgezeigt werden. Wir haben fünf wesentliche Kategorien der KI-Nutzung identifiziert. Diese Kategorien spiegeln wider, auf welche Weise KI in ein Unternehmen Einzug hält und dort funktioniert: über Lieferanten, Entwicklungspipelines, Betriebssysteme, Benutzerinteraktionen und Governance-Prozesse.

Zusammenfassend lauten die fünf Kategorien:



Die obige Tabelle bietet einen allgemeinen Überblick über diese Bereiche und zeigt auf, wie die Risiken den wichtigsten Sicherheitsfunktionen über den gesamten KI-Lebenszyklus hinweg zugeordnet sind. Sie bietet einen kurzen Überblick darüber, worauf sich die Abwehrmaßnahmen konzentrieren sollten, bevor auf den folgenden Seiten die einzelnen Fähigkeiten näher beleuchtet werden.

Schutz vor Missbrauch und unerwarteten Verhaltensweisen

Fähigkeiten

Analyse bössartiger Eingabeaufforderungen

Die Erkennung und Analyse schädlicher oder manipulativer Eingabeaufforderungen, die darauf abzielen, KI-Systeme oder -Agenten zu unerwünschten oder unsicheren Handlungen zu zwingen.

Risikobeschreibung: KI-Agenten können dazu veranlasst oder dazu verleitet werden, unter Ausnutzung ihrer Berechtigungen und ihres Datenzugriffs unerwünschte Aktionen durchzuführen. Dies kann schwer zu überwachen oder zu blockieren sein, insbesondere bei externen KI-Agenten. Verfügt ein Bedrohungsakteur über ausreichende Zugriffsrechte, kann er ein KI-Modell untersuchen und Daten daraus extrahieren, die eigentlich darin verborgen bleiben sollten.

Sicherheitsfunktion: Erkennung und, soweit möglich, Blockierung von Anfragen mit böswilliger Absicht, unabhängig davon, ob diese intern oder extern, direkt oder indirekt erfolgen. Überwachen Sie Versuche, Modell-daten zu extrahieren oder zu manipulieren, um Missbrauch zu verhindern und sensible Informationen im Rahmen von KI-Interaktionen zu schützen.

Unbeabsichtigte Anwendungsfälle

Der Missbrauch oder die Umnutzung von KI-Systemen in einer Weise, die ursprünglich nicht vorgesehen oder genehmigt war, was häufig zu unerwarteten Verhaltensweisen oder Risiken führt.

Risikobeschreibung: KI-Agenten und -Systeme, die missbraucht werden, können eine Vielzahl unerwünschter Auswirkungen haben. Die Risiken werden dadurch verstärkt, dass es schwierig ist, einen Missbrauch durch autorisierte Parteien zu erkennen, wenn die Schutzmaßnahmen nicht wirksam sind.

Sicherheitsfunktion: Unbefugte oder nicht genehmigte Nutzungen von KI-Systemen erkennen und blockieren. Überwachen Sie unbeabsichtigte Datenoffenlegungen oder Datenmissbrauch, um sicherzustellen, dass KI-Modelle ausschließlich innerhalb genehmigter und konformer Parameter betrieben werden.

Zugriff auf den Prompt-Verlauf

Die Speicherung und Verwaltung von KI-Prompt-Historien, die sensible Daten, Kontextinformationen oder Verhaltensmuster enthalten können.

Risikobeschreibung: Ein unbefugter Zugriff auf den Verlauf von Eingabeaufforderungen kann vertrauliche Informationen offenlegen, selbst wenn der Dienst, an den die Eingabeaufforderungen gesendet wurden, sicher ist und die Informationen nicht speichert.

Sicherheitsfunktion: Beschränkung des Zugriffs und Überwachung des Systems auf Sicherheitsverletzungen, da ein Bedrohungsakteur, der die Kontrolle über das Konto erlangt hat, diese Informationen für weitere Angriffe ausnutzen könnte.

Böswillige und halluzinatorische Ausgabeeffekte

Die Erzeugung von KI-Ergebnissen, die ungenau, irreführend oder absichtlich schädlich sind und möglicherweise das Handeln von Menschen beeinflussen.

Risikobeschreibung: Abgesehen von Risiken im Zusammenhang mit der Automatisierung unter Verwendung halluzinierter oder böswilliger Ausgaben (wie beispielsweise „Package-Squatting“ und unerwartete Aktionen von KI-Agenten) können diese Ausgaben auch dazu führen, dass Menschen unerwünschte Handlungen vornehmen. Der Umfang dieser potenziellen Maßnahmen ist weitaus größer als der automatisierter Maßnahmen, beispielsweise wenn sie angewiesen werden, Geld zu überweisen oder böswillige Einstellungen an einem kritischen IT-System vorzunehmen, auf das KI-Agenten bewusst keinen Zugriff haben.

Sicherheitsfunktion: Überwachung auf schädliche oder gefälschte KI-Ergebnisse, die Benutzer irreführen oder unbeabsichtigte Aktionen auslösen könnten. Einführung einer Halluzinationserkennung und einer Ausgabevalidierung, um Fehlinformationen oder unsicheres Verhalten zu verhindern, das durch KI-generierte Inhalte beeinflusst wird.

Überwachung und Steuerung von KI im Betrieb

Fähigkeiten

Aktionen und Verbindungen von KI-Agenten

Für Unternehmen, die KI-Tools entwickeln

Die Überwachung des Echtzeitverhaltens von KI-Agenten, einschließlich ausgehender Verbindungen, Datenaustausch und Interaktionen mit anderen Agenten oder Diensten.

Risikobeschreibung: KI-Agenten könnten ihre Berechtigungen missbräuchlich nutzen oder auf Daten oder Aktionen zugreifen, die nicht vorgesehen waren. Außerdem können sie beim Aufruf anderer KI-Agenten oder Dienste (einschließlich MCP-Servern) Daten preisgeben oder unerwünschte Aktionen auslösen. KI-Agenten verfügen nicht über die Vorsicht menschlicher Mitarbeiter, wenn die Sicherheitsvorkehrungen mögliche Probleme nicht vorhersehen oder versagen.

Sicherheitsfunktion: Überwachung von KI-Agenten in Echtzeit hinsichtlich ausgehender Verbindungen, Anfragen und Befehlen an externe Dienste (einschließlich MCP-Server) oder andere Agenten. Verfolgung der Nutzung von Zugangsdaten und gegebenenfalls Durchsetzung von Sperren, um Datenlecks oder unbefugte Handlungen zu verhindern.

KI-Gateway

Eine zentralisierte Schnittstelle oder Steuerungsebene, die die Interaktionen von KI-Systemen verwaltet, darunter die Übermittlung von Eingabeaufforderungen, die Weiterleitung an Modelle, die Kostenkontrolle und die Durchsetzung des Datenschutzes.

Risikobeschreibung: Der Zugriff auf KI-Systeme über Gateways ist möglicherweise nicht möglich, oder ein Gateway ist falsch konfiguriert oder fällt aus. Dies kann zu mangelnder Sicherheit, fehlerhafter Modellführung, unkontrollierten Kosten oder unzureichender Überwachung und Echtzeitsteuerung führen.

Sicherheitsfunktion: Bereitstellung sowie zentrale Verwaltung von KI-Gateways und Durchsetzung von deren Nutzung, wo immer möglich. Einbeziehen von Protokollen, Warnmeldungen und Metriken in die System- und Sicherheitsüberwachung.

Wenn KI-Systeme gezielt eingesetzt werden, um Daten lokal (beispielsweise vor Ort) zu speichern, sollten die Sicherheitsmaßnahmen entsprechend angepasst sein.

KI-Agenten-Zustandserfassung

Für Unternehmen, die KI-Tools entwickeln

Der Prozess der Sicherung des Betriebszustands und der Informationen potenziell vorübergehender oder kurzlebiger KI-Komponenten, insbesondere in Cloud-Umgebungen, um Untersuchungen und Analysen nach einem Vorfall zu ermöglichen.

Risikobeschreibung: Wenn eine kurzlebige KI-Agent-Komponente (wie sie in Cloud-basierten Systemen häufig vorkommt) eine Aktion ausführt, die untersucht oder dokumentiert werden muss, kann es vorkommen, dass die Komponente verschwindet, bevor Sicherheitsprozesse sie erfassen können. Das Sicherheitsteam ist möglicherweise gar nicht in der Lage, bestimmte Komponenten zu erfassen und zu sichern.

Sicherheitsfunktion: Automatische Erfassung und Speicherung des vollständigen Zustands eines verdächtigen KI-Agenten, einschließlich temporärer Cloud-Instanzen. Dies ermöglicht Ermittlungen und Gegenmaßnahmen und stellt sicher, dass Beweismittel und der Kontext für die Analyse erhalten bleiben.

tark regulierte Branchen sollten diesem Risiko besondere Aufmerksamkeit widmen

KI-Datenübertragung

Der Datenaustausch zwischen KI-Komponenten, Systemen oder externen Endpunkten, einschließlich sowohl automatisierter als auch von Menschen initiiertem Datenaustausch.

Risikobeschreibung: KI-Systeme übermittelte Daten können gegen die Datenhoheit oder gesetzliche Vorschriften (wie die DSGVO) verstoßen. Daten können von Menschen, von automatisierten Systemen oder durch Entscheidungen von KI-Agenten übermittelt werden. Datenübertragungen an oder von KI können schwache Kryptografie verwenden und müssen in den PQC-Übergangsplänen berücksichtigt werden.

Sicherheitsfunktion: Sicherung von Daten während der Übertragung zwischen KI-Komponenten mithilfe von Verschlüsselungsverfahren, gegebenenfalls einschließlich postquantenkryptografischer Verfahren. Sicherstellen, dass alle Datenübermittlungen den einschlägigen Vorschriften zur Datenübermittlung und zum Datenschutz, wie beispielsweise der DSGVO, entsprechen.

Schutz der KI-Entwicklung und -Infrastruktur

Fähigkeiten

KI-Sicherheitskonfiguration

Für Unternehmen, die KI-Tools entwickeln

Die Einrichtung und laufende Verwaltung von Sicherheitsparametern, Schutzmaßnahmen und Zugriffskontrollen, die die Funktionsweise von KI-Komponenten regeln.

Risikobeschreibung: Konfigurationen, die KI-Agenten sowie zugehörigen und unterstützenden Einzelkomponenten zugewiesen werden, können fehlerhaft sein und unbefugten Zugriff ermöglichen oder eine schwache Sicherheit gegen Angriffe aufweisen. Eine Fehlkonfiguration kann auch vorgesehene Schutzmechanismen aufheben.

Sicherheitsfunktion: Gewährleistung sicherer Konfigurationen für alle KI-Komponenten, einschließlich Schutzmaßnahmen, Richtlinien und Berechtigungen. Eine automatisierte Überwachung und Analyse sowie regelmäßige Überprüfungen der Richtlinien und deren Durchsetzung tragen dazu bei, ein robustes und konformes KI-Sicherheitskonzept aufrechtzuerhalten.

KI-Agenten-Architekturen

Für Unternehmen, die KI-Tools entwickeln

Das Design und die Struktur von KI-Agenten und ihren miteinander verbundenen Komponenten, die bestimmen, wie sie interagieren, Daten austauschen und Sicherheitsgrenzen durchsetzen.

Risikobeschreibung: Die aus mehreren Komponenten bestehende Architektur eines eingesetzten KI-Agenten oder von KI-Agenten, die in einem System zusammenarbeiten, kann insgesamt Mängel aufweisen und unberechtigten Zugriff ermöglichen oder keine Schutzmaßnahmen gegen unbeabsichtigte Nutzung bieten – unabhängig von Problemen mit einzelnen Komponenten.

Sicherheitsfunktion: Durchführung einer statischen Analyse von KI-Agentenarchitekturen zur Identifizierung von Sicherheitslücken. Bewertung lokaler Netzwerkexpositionen, Anmeldedaten für Dienste und Berechtigungen, um unbefugten Zugriff oder laterale Bewegung zu verhindern. Eine ordnungsgemäße Segmentierung und eine ordnungsgemäße Verwaltung von Zugangsdaten sind entscheidend für die Reduzierung architektonischer Risiken.

Entwicklungsteams können auf Unterstützungsdienste für die Erstellung von KI-Agenten zurückgreifen. So weit wie möglich auf einen einzigen Dienst vereinheitlichen, der akzeptable Sicherheitsfunktionen bietet.

IaC-Scanning

Für Unternehmen, die KI-Tools entwickeln

Die automatisierte Analyse von „Infrastructure-as-Code“-Vorlagen zur Erkennung von Fehlkonfigurationen oder Sicherheitslücken vor der Bereitstellung.

Risikobeschreibung: Bei der Nutzung von „Infrastructure-as-Code“ können Probleme bei der KI-Sicherheitskonfiguration und der KI-Agentenarchitektur auftreten, die auf unzulässigen Zugriff auf den Code, das Einschleusen von Schadcode aus internen oder externen Quellen oder auf Fehler zurückzuführen sind.

Sicherheitsfunktion: Verwendung spezieller Analyse-Tools, um Infrastructure-as-Code (IaC) auf Fehlkonfigurationen oder verborgene Risiken zu überprüfen. Die frühzeitige Erkennung von Sicherheitsproblemen in Bereitstellungsskripten trägt dazu bei, zu verhindern, dass sich Schwachstellen in KI-Umgebungen ausbreiten.

Überprüfung von KI-generiertem Code

Für Unternehmen, die KI-Tools entwickeln

Die Überprüfung von Code, der von KI-Systemen (Coding-Copilots) erstellt wurde, um Sicherheitslücken, böswillige Einfügungen oder unzuverlässige Abhängigkeiten aufzudecken.

Risikobeschreibung: Generierter Code kann Sicherheitslücken aufweisen, sei es versehentlich oder aufgrund einer böswilligen Wirkung des generierenden Modells. Dazu kann das Hinzufügen bösartiger Abhängigkeiten gehören; so erstellen Bedrohungsakteure beispielsweise Pakete mit erfundenen Namen. Generierter Code kann deutlich schwerer zu verstehen und zu debuggen sein als manuell geschriebener Code. Probleme treten eher auf, wenn der Umfang und die Komplexität des generierten Codes zunehmen.

Sicherheitsfunktion: Überprüfung von KI-generiertem Code auf Fehler, unsichere Logik oder böswillige Einfügungen. Erkennung von Problemen wie fiktiven Abhängigkeiten oder „Package Squatting“, um zu verhindern, dass kompromittierter oder unsicherer Code in Produktionsumgebungen gelangt.

Sicherung der KI-Lieferkette

Fähigkeiten

Einsatz von Shadow AI

Die Nutzung nicht genehmigter oder nicht zugelassener KI-Tools durch Mitarbeiter oder Abteilungen, häufig außerhalb der offiziellen Sicherheits- und Compliance-Kontrollen. Diese Dienste werden in der Regel extern betrieben und können ohne Aufsicht sensible Geschäftsdaten erfassen oder für Trainingszwecke nutzen.

Risikobeschreibung: Daten gelangen an nicht autorisierte externe Parteien, darunter möglicherweise geschützte Informationen, Finanzdaten, Kundendaten oder vertrauliches geistiges Eigentum. Nicht zugelassene Dienste nutzen die übermittelten Daten wahrscheinlich zu Trainingszwecken.

Sicherheitsfunktion: Erkennung und Blockierung der Nutzung nicht genehmigter KI-Dienste, wie beispielsweise nicht genehmigter generativer KI-Tools oder der Nutzung nicht vertrauenswürdiger externer MCP-Server durch genehmigte Tools. Eine unüberwachte Nutzung kann sensible Daten gefährden und Lücken in der Sicherheitsüberwachung entstehen lassen.

KI-Anbieter, die Agenten direkt bereitstellen

Externe KI-Agenten oder GenAI-Dienste, die in Geschäftssysteme integriert sind und internen Zugriff oder Berechtigungen erhalten, um automatisierte Aufgaben im Namen von Benutzern oder Anwendungen auszuführen.

Risikobeschreibung: Die Aktivitäten dieser Agenten sind nach ihrer Auslösung möglicherweise weder auf der Befehlsebene noch auf der Verbindungsebene sichtbar oder kontrollierbar.

Sicherheitsfunktion: Überwachung externer KI-Agenten, denen interner Zugriff gewährt wurde, da diese häufig mit weitreichenden Berechtigungen und begrenzter Transparenz arbeiten. Durchsetzen strenger Zugriffskontrollen, Protokollierungen und Überwachungsmaßnahmen, um Missbrauch oder unbeabsichtigte Aktionen innerhalb interner Umgebungen zu verhindern.

KI im Lieferantenmanagement

Der Einsatz externer Anbieter oder Dienste von Drittanbietern, die KI zur Verarbeitung oder Analyse von Geschäftsdaten einsetzen. Dies umfasst sowohl ausdrücklich als KI-Lösungen bezeichnete Angebote als auch Fälle, in denen Anbieter KI in ihre Produkte oder Arbeitsabläufe integrieren, ohne dies den Kunden ausdrücklich mitzuteilen.

Risikobeschreibung: Anbieter, die einen bekannten KI-Dienst bereitstellen, sichern Geschäftsdaten möglicherweise nicht angemessen – weder vertraglich noch aufgrund betrieblicher Mängel. Lieferanten können KI intern und unbemerkt einsetzen und so der Kontrolle durch das Unternehmen entgehen, das ihnen vertrauliche Daten zur Verfügung stellt.

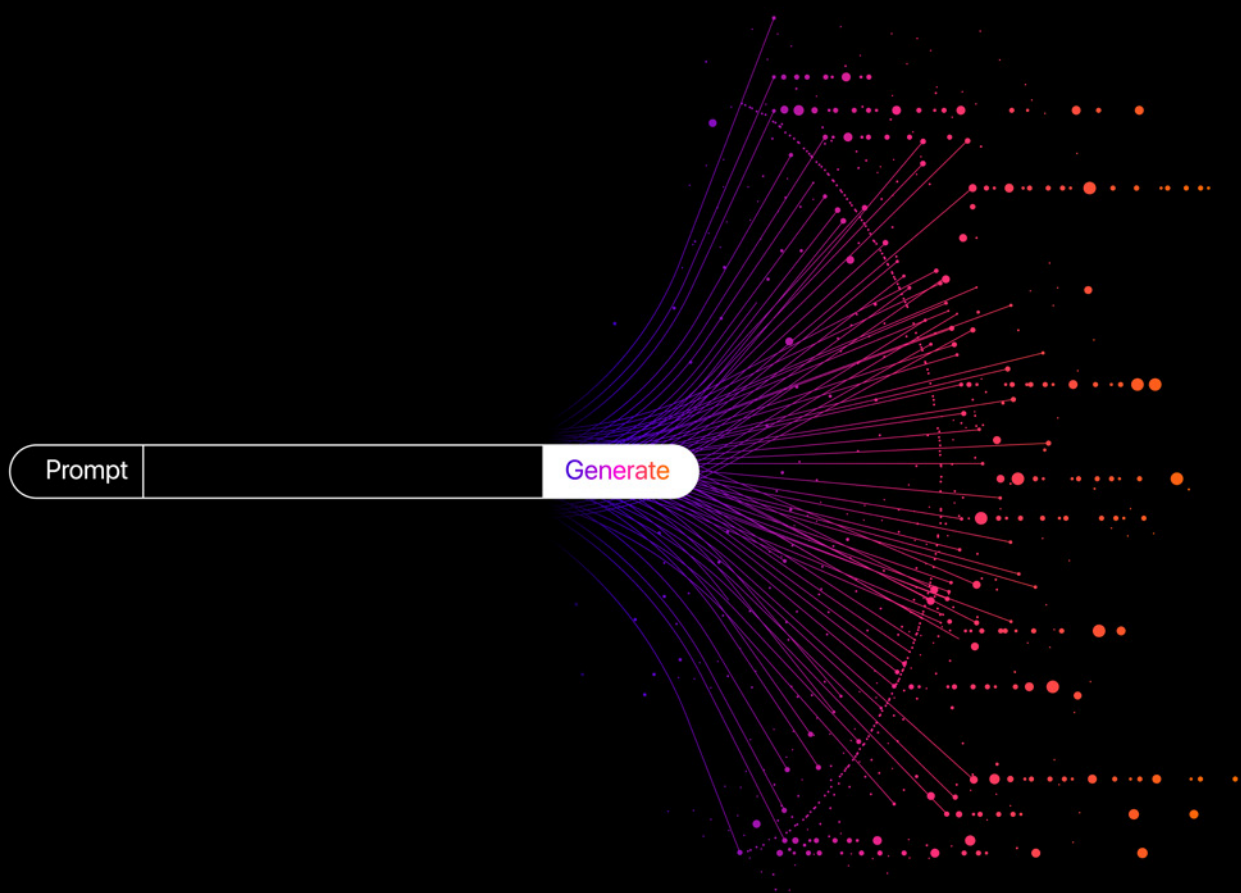
Sicherheitsfunktion: Bewertung von Anbietern, die KI auf Ihre Geschäftsdaten anwenden (unabhängig davon, ob diese offengelegt oder verborgen sind). Überprüfen, wie externe KI-Systeme diese Daten verarbeiten, speichern und schützen, um eine unbefugte Nutzung oder unbeabsichtigte Offenlegung durch Dienste von Drittanbietern zu verhindern.

Code- und Abhängigkeitsprüfung für KI-Agenten

Der Prozess der Überprüfung des Quellcodes von KI-Agenten und der zugehörigen Softwareabhängigkeiten, um Schwachstellen, Lizenzprobleme oder böswillige Einfügungen zu identifizieren. Dadurch wird sichergestellt, dass die Komponenten, die zum Aufbau oder zur Erweiterung von KI-Systemen verwendet werden, sicher und konform sind und wie vorgesehen funktionieren.

Risikobeschreibung: Software-Abhängigkeiten enthalten möglicherweise ungeeignete Lizenzen oder wurden kompromittiert, um bösartig zu sein – unabhängig davon, ob sie Open-Source oder kommerziell sind. Für KI-Agenten entwickelter Code kann Sicherheitslücken aufweisen, die die beabsichtigten Sicherheitsziele beeinträchtigen oder aufheben.

Sicherheitsfunktion: Führen Sie eine statische Sicherheitsanalyse des Codes interner KI-Agenten durch, um Schwachstellen und unsichere Konfigurationen zu identifizieren. Überprüfen Sie alle Softwareabhängigkeiten, um bekannte Schwachstellen oder bösartige Inhalte vor der Bereitstellung zu erkennen. Regelmäßige Scans helfen, die Code-Integrität aufrechtzuerhalten und Risiken in der KI-Entwicklungspipeline zu reduzieren.



Angebot an KI-Modellen

Für Unternehmen, die KI-Tools entwickeln

Der Prozess der Beschaffung oder Integration externer KI-Modelle, ob Open-Source, von Drittanbietern oder von Herstellern bereitgestellt, in die Systeme oder Entwicklungspipelines einer Organisation.

Risikobeschreibung: Aus externen Quellen importierte Modelle können schädliche oder fehlerhafte Inhalte und Anweisungen enthalten, die zu Ergebnissen führen, welche KI-Agenten oder Nutzer dazu veranlassen, unerwünschte Handlungen auszuführen. Dazu könnten das Ausmaß der Verzerrung, die Genauigkeit, die Abweichung und andere Größen gehören, die für das Unternehmen oder dessen spezifische Anwendungsfälle inakzeptabel sind. Möglicherweise verfügen sie nicht über Lizenzen, die für den Einsatz im Unternehmen geeignet sind.

Sicherheitsfunktion: Die Absicherung der Modell-Lieferkette bedeutet, sowohl die Herkunft als auch den Inhalt zu überprüfen. Jedes externe Modell sollte auf darin enthaltene Risiken überprüft, auf ordnungsgemäße Lizenzierung hin bestätigt und unter Verwendung einer Versionskontrolle sicher gespeichert werden.

KI-Trainingsdaten

Für Unternehmen, die KI-Tools entwickeln

Die Erfassung und Aufbereitung von Datensätzen, die zum Trainieren oder Feinabstimmen von KI-Modellen oder im Rahmen von TEVV (Test, Bewertung, Validierung und Verifizierung) verwendet werden. Diese Datensätze können aus internen Quellen, öffentlichen Repositories oder von Drittanbietern stammen und beeinflussen direkt das Modellverhalten, die Genauigkeit und die Compliance.

Risikobeschreibung: Trainingsdaten können bösartige Daten enthalten, die dazu führen, dass ein trainiertes Modell unerwünschte Auswirkungen hat. Dazu könnten unrechtmäßig geschützte Daten oder Daten gehören, für die keine ordnungsgemäße Lizenz vorliegt, oder es kann sein, dass ein Datensatz insgesamt nicht ordnungsgemäß für die Nutzung im Unternehmen lizenziert ist.

Sicherheitsfunktion: Die Sicherung von Trainingsdaten beginnt damit, genau zu wissen, woher sie stammen und was sie enthalten. Überprüfen Sie vor der Verwendung die Herkunft, die Lizenzierung und die Integrität der Daten. Schützen Sie sich vor interner oder externer Datenverfälschung. Erkennen und bewältigen Sie Risiken im Zusammenhang mit der Datenlizenzierung, insbesondere bei Grundmodellen, sowie die Möglichkeit, dass Trainingsdaten direkt in den Modellausgaben wiedergegeben werden.

Stärkung der Einsatzbereitschaft und Aufsicht

Fähigkeiten

Berichterstattung

Der Prozess der Erfassung und Kommunikation von KI-bezogenen Risiken, Kennzahlen und Trends zur Unterstützung einer fundierten Entscheidungsfindung und Governance.

Risikobeschreibung: Ziele und Kennzahlen sind für ein effektives Management und eine kontinuierliche Verbesserung von entscheidender Bedeutung. Ohne ein effektives KI-Management-Reporting können kritische Probleme oder Trends übersehen werden. Das Unternehmen als Ganzes könnte unerwünschte Entscheidungen in Bezug auf KI treffen, ohne dass die Interessen der Betroffenen angemessen berücksichtigt werden.

Sicherheitsfunktion: Erstellen übersichtlicher Zusammenfassungen von KI-bezogenen Sicherheitsereignissen, Risikokennzahlen und der Systemleistung. Nachverfolgung der Umsetzung der geplanten KI-Anwendungsfälle und Meldung etwaiger unbeabsichtigter oder unbefugter Nutzungen, um die Aufsicht und Steuerung zu gewährleisten.

KI-Penetrationstests, Red- und Purple-Teaming

Das Testen und Validieren von live betriebenen KI-Systemen durch simulierte Angriffe und adversarische Übungen, um Schwachstellen und Lücken in der Erkennung aufzudecken.

Risikobeschreibung: In Live-Systemen treten häufig latente Probleme auf, die weder in der Konfiguration noch im Code erkennbar sind. Ohne Penetrationstests könnten Bedrohungsakteure bestimmte Schwachstellen entdecken, die den Sicherheitsteams nicht bekannt sind. Ohne Tests durch Red- oder Purple-Teams könnten umfassendere Schwachstellen bei den Erkennungs- und Aufdeckungsfähigkeiten vorhanden sein, die jedoch unbekannt sind.

Sicherheitsfunktion: Durchführung von Penetrationstests und Red- oder Purple-Team-Übungen zur Simulation realer Angriffe auf live betriebene KI-Systeme. Prüfung auf ausnutzbare Schwachstellen, insbesondere in nach außen gerichteten Umgebungen, um Angreifertechniken zu antizipieren und deren Auswirkungen zu mindern, bevor sie zum Tragen kommen.

Schulungen und Übungen für Sicherheitsteams

Fortlaufende Schulungen und Simulationsübungen, die darauf abzielen, in den Sicherheitsteams KI-spezifische Kompetenzen in den Bereichen Erkennung, Untersuchung und Reaktion auf Vorfälle aufzubauen.

Risikobeschreibung: Unzureichende individuelle Schulungen oder mangelnde Incident-Response-Übungen im Team können dazu führen, dass Vorfälle übersehen werden, Untersuchungen unzureichend durchgeführt werden oder es nicht gelingt, ein Cybersicherheitsproblem oder einen Bedrohungsakteur zu stoppen oder zu beseitigen.

Sicherheitsfunktion: Entwicklung KI-spezifischer Kompetenzen und Vorbereitung der Sicherheitsteams auf den Einsatz von KI. Regelmäßige Schulungen und Simulationen stärken die Kompetenz bei der Erkennung von KI-bezogenen Vorfällen, der Reaktion darauf und der Wiederherstellung nach solchen Vorfällen.



Ein neues Verständnis von KI-Sicherheit für die Unternehmensleitung

Dieselbe Intelligenz, die geschäftliche Entscheidungen beschleunigt, kann – wenn sie ungeschützt bleibt – Risiken in beispiellosem Tempo und Ausmaß verstärken.

Für Sicherheitsverantwortliche besteht die Aufgabe nicht nur darin, KI-Systeme zu schützen, sondern auch darin, deren Komplexität in klare, strategische Leitlinien für die Führungsebene zu übersetzen.

Auf Vorstandsebene kann dies damit beginnen, dass das Thema KI-Sicherheit unter den Gesichtspunkten Vertrauen, Verantwortlichkeit und Widerstandsfähigkeit betrachtet wird:

Vertrauen

Vorstände wollen die Gewissheit, dass KI-gestützte Entscheidungen korrekt, nachvollziehbar, datenschutzkonform und vor Manipulationen geschützt sind. CISOs sollten die Kontrollmaßnahmen hervorheben, mit denen die Datenintegrität, das Modellverhalten und die Herkunft jeder Komponente in der KI-Lieferkette überprüft werden.

Rechenschaftspflicht

Stellen Sie klar, wer für die einzelnen Ebenen des KI-Risikos verantwortlich ist: Datenwissenschaftler, Cloud-Anbieter, Lieferanten und Sicherheitsteams haben jeweils festgelegte Zuständigkeiten. Vorstände reagieren auf klare Strukturen und gemeinsame Verantwortung, nicht auf Unklarheiten.

Stärken Sie die Rechenschaftspflicht, indem Sie die Sicherheitsgrundsätze für KI mit etablierten Governance-Rahmenwerken wie dem NIST AI Risk Management Framework (RMF), ISO/IEC 42001 sowie den Leitlinien der CISA und des NCSC verknüpfen. Diese Verknüpfung zeugt von Reife, gewährleistet die einheitliche Einhaltung aller Compliance-Vorgaben und bietet einen klaren Maßstab für die Aufsicht.

Resilienz

Betonen Sie, dass KI mittlerweile Teil der kritischen Infrastruktur des Unternehmens ist. Sie zu schützen bedeutet, Kontinuität im Falle eines Angriffs, Nachvollziehbarkeit bei Prüfungen und Flexibilität bei Änderungen der regulatorischen Rahmenbedingungen zu gewährleisten.

Da KI zunehmend in alle Geschäftsabläufe integriert wird, wird es für die Widerstandsfähigkeit von Unternehmen entscheidend sein, ihre Sicherheit mit Zielen wie dem Schutz geistigen Eigentums, der Betriebssicherheit und dem Vertrauen der Stakeholder zu verknüpfen. Unternehmen, die über Managementsysteme verfügen, die den Ausbau von KI-Anwendungsfällen aktiv fördern und gleichzeitig dessen Sicherheit gewährleisten, sind am besten aufgestellt, um sicher innovativ zu sein, sich schnell anzupassen und das Vertrauen von Kunden, Aufsichtsbehörden und Investoren zu bewahren.

Fazit

Dieselben Kräfte, die KI so transformativ machen, erschweren es zugleich, ihre Sicherheit zu gewährleisten, wenn sie in SaaS-Tools, Cloud-Infrastrukturen, Anbietersystemen und selbst entwickelten Anwendungen integriert ist. In einem solchen Umfeld können Transparenz, Eigenverantwortung und Rechenschaftspflicht schnell verschwinden.

Um KI sicher zu machen, muss diese Klarheit wiederhergestellt werden. Dazu muss man nicht nur verstehen, wie KI aufgebaut ist, sondern auch, wo sie zum Einsatz kommt, wer mit ihr interagiert und wie ihre Entscheidungen das Geschäft beeinflussen. Indem sie die Sicherheitsrisiken im Zusammenhang mit KI aus dieser breiteren Perspektive betrachten, die Lieferketten, Entwicklung, Echtzeitbetrieb und Governance umfasst, können Unternehmen beginnen, Unsicherheit in Sicherheit umzuwandeln.

Der Weg zu sicherer KI ist noch nicht zu Ende, doch er beginnt mit dem Bewusstsein: zu wissen, wo KI zum Einsatz kommt, wie sie sich verhält und wie man verantwortungsvoll damit umgeht.

KI sicher zu entwickeln bedeutet, sie verantwortungsbewusst zu entwickeln.

Erfahren Sie in unserem Beitrag „Auf dem Weg zu verantwortungsvoller KI in der Cybersicherheit“, wie sich Innovation und Governance in Einklang bringen lassen

Erfahren Sie mehr ➔



Entdecken Sie die Vielfalt der KI-Typen in der Cybersicherheit

Erfahren Sie mehr über die Technologien, die hinter moderner Resilienz stehen – vom überwachten maschinellen Lernen bis hin zu NLP – und wie diese zusammenwirken, um neue Bedrohungen abzuwehren.

Erfahren Sie mehr ➔

■ Über Darktrace

Darktrace ist ein weltweit führender Anbieter von KI-gestützter Cybersicherheit, der Unternehmen dabei hilft, der sich ständig verändernden Bedrohungslandschaft einen Schritt voraus zu sein. Darktrace wurde 2013 in Cambridge, Großbritannien, gegründet und bietet eine unverzichtbare Cybersicherheitsplattform, die Unternehmen vor unbekannten Bedrohungen schützt, indem sie künstliche Intelligenz einsetzt, die in Echtzeit von jedem einzelnen Unternehmen lernt. Die Plattform und die Dienstleistungen von Darktrace werden von über 2.700 Mitarbeitern unterstützt, die weltweit fast 10.000 Kunden schützen. Mehr dazu erfahren Sie unter www.darktrace.com.