

Glossar

Ausgabe 202611

Copyright © EXIN Holding B.V. 2026. All rights reserved.
EXIN® is a registered trademark of EXIN.

No part of this publication may be reproduced, stored, utilized or transmitted in any form or by any means, electronic, mechanical, or otherwise, without the prior written permission from EXIN.

Einführung

Dies ist das Glossar für die Zertifizierung EXIN AI Security Professional based on the OWASP AI Exchange (AISP.DE).

Es enthält Definitionen der grundlegenden Begriffe, die im Vorbereitungshandbuch aufgeführt sind und in der Prüfung verwendet werden, um die Kandidatinnen und Kandidaten bei ihrer Prüfungsvorbereitung zu unterstützen.

Die bereitgestellten Definitionen sind auf die OWASP AI Exchange abgestimmt.

Glossar

Die folgenden Begriffe sind nach dem englischen Begriff alphabetisch geordnet und zusammen mit ihren Definitionen aufgeführt.

Begriff	Definition
Verantwortung *accountability*	Der Grundsatz, wonach Einzelpersonen oder Organisationen für ihre Handlungen und Entscheidungen Rechenschaft ablegen müssen und aufgefordert werden können, Ergebnisse zu erläutern, zu begründen oder die Verantwortung dafür zu übernehmen.
adversarialer Angriff *adversarial attack*	Mehrdeutiger Begriff, der manchmal für jeden beliebigen Angriff verwendet wird, manchmal aber auch für den Versuch, Eingabedaten, Daten oder ein Modell absichtlich so zu gestalten oder zu manipulieren, dass ein System des maschinellen Lernens falsche, schädliche oder vom Angreifer gewählte Ergebnisse liefert.
adversariales Testen *adversarial testing*	Tests durch die Durchführung adversarialer Angriffe, um ausnutzbare Schwachstellen aufzudecken, bevor echte Angreifer dies tun.
agentische KI *agentic AI*	Systeme, die auf KI-Agenten basieren, die planen, Maßnahmen ergreifen und mit Tools, Daten oder anderen Agenten interagieren können, um Ziele mit begrenztem menschlichem Eingriff zu erreichen.
KI (künstliche Intelligenz) *AI (artificial intelligence)*	Die Fähigkeit eines Computersystems, Aufgaben auszuführen, die typischerweise menschenähnliche Intelligenz erfordern, wie beispielsweise Schlussfolgern, Lernen, Mustererkennung oder das Erstellen von Inhalten, sei es durch regelbasierte Methoden oder durch maschinelles Lernen.
KI-Agent *AI agent*	Eine Softwarekomponente, die autonom über Maßnahmen entscheiden und diese ausführen kann (beispielsweise das Aufrufen von Funktionen oder Tools), um ein Ziel zu erreichen, anstatt lediglich eine Antwort zu generieren.
KI-Assets *AI assets*	Die KI-spezifischen Elemente eines bestimmten Systems oder einer bestimmten Organisation, die geschützt werden müssen, wie beispielsweise Trainingsdaten und Modellparameter.
KI-Jailbreak *AI jailbreak*	Ein Angriff durch Prompt-Injection, der die integrierten Sicherheits- oder Inhaltsbeschränkungen eines Modells umgeht, um es dazu zu bringen, eine Ausgabe zu erzeugen, die es eigentlich ablehnen sollte.

Begriff	Definition
KI-Managementsystem (AIMS) *AI management system (AIMS)*	Ein organisatorisches Rahmenwerk aus Richtlinien und Prozessen, wie es beispielsweise in der Norm ISO/IEC 42001 definiert ist, zur verantwortungsvollen und sicheren Steuerung von KI.
KI-Modell *AI model*	Ein mathematisches oder rechnerisches Konstrukt, das Eingaben auf Ausgaben abbildet und entweder aus Daten gelernt oder durch Regeln definiert wurde.
Anbieter eines KI-Modells *AI model provider*	Eine Organisation oder Partei, die ein KI-Modell entwickelt, trainiert oder bereitstellt, damit andere es nutzen oder darauf aufbauen können.
Datenschutzgrundsätze für KI *AI privacy principles*	Eine Reihe von Grundsätzen für den Umgang mit personenbezogenen Daten, die Aspekte wie Genauigkeit, Einwilligung, Datenminimierung, Erklärbarkeit, Fairness, Transparenz und Speicherbegrenzung abdecken.
• **Genauigkeit** *accuracy*	Im Zusammenhang mit personenbezogenen Daten: der Grundsatz, dass personenbezogene Daten korrekt sein und, soweit erforderlich, auf dem neuesten Stand gehalten werden sollten. Im Zusammenhang mit Modellen: das Ausmaß, in dem die Ergebnisse eines KI-Systems korrekt sind.
• **Einwilligung** *consent*	Die freiwillig erteilte, spezifische, in voller Kenntnis der Sachlage erfolgte und unmissverständliche Zustimmung einer betroffenen Person zur Verarbeitung ihrer personenbezogenen Daten.
• **Datenminimierung und Speicherbegrenzung** *data minimization and storage limitation*	Der Grundsatz, nur die für einen bestimmten Zweck erforderlichen Daten zu erheben und diese nicht länger aufzubewahren, als es für diesen Zweck erforderlich ist.
• **Rechte der betroffenen Person** *data subject rights*	Individuelle Rechte in Bezug auf personenbezogene Daten, wie beispielsweise das Recht auf Auskunft, Berichtigung und Löschung.
• **Erklärbarkeit** *explainability*	Der Grundsatz, dass betroffene Personen darüber informiert werden sollten, wie ihre Daten verwendet werden, und – sofern automatisierte Entscheidungen sie betreffen – nachvollziehen können sollten, durch welche Schritte diese Entscheidungen zustande gekommen sind.
• **Fairness und Rechtmäßigkeit** *fairness and lawfulness*	Der Grundsatz, dass personenbezogene Daten auf rechtlich gerechtfertigte Weise verarbeitet werden müssen und Einzelpersonen oder Gruppen dadurch nicht ungerechtfertigt benachteiligt werden dürfen.

Begriff	Definition
• **Transparenz** *transparency*	Der Grundsatz, dass Einzelpersonen darüber informiert werden sollten, wie ihre Daten erhoben, genutzt, weitergegeben und gespeichert werden, und dass sie, sofern Entscheidungen durch Automatisierung getroffen werden, in der Lage sein sollten, die Vorgehensweise nachzuvollziehen, die zu diesen Entscheidungen geführt hat.
• **Verwendungsbegrenzung und Zweckbestimmung** *use limitation and purpose specification*	Der Datenschutzgrundsatz, wonach personenbezogene Daten, die für einen bestimmten Zweck erhoben wurden, nicht ohne ordnungsgemäße Rechtsgrundlage einfach wiederverwendet werden dürfen, um ein KI-Modell für einen anderen, unvereinbaren Zweck zu trainieren.
KI-Red-Teaming *AI red teaming*	Eine strukturierte, adversariale Testübung, bei der Tester wie echte Angreifer vorgehen, um Schwachstellen in den Bereichen Sicherheit, Safety oder Ethik in einem System aufzudecken.
KI-Sicherheit *AI security*	Die Disziplin zum Schutz von KI-Systemen, deren Daten und der zugrunde liegenden Infrastruktur vor Sicherheitsbedrohungen, die KI-spezifische Eigenschaften ausnutzen, sowie vor den herkömmlichen Cybersicherheitsrisiken.
Governance der KI-Sicherheit *AI security governance*	Die organisatorischen Strukturen, Richtlinien und Prozesse der Aufsicht, die sicherstellen, dass KI-bezogene Sicherheitsrisiken identifiziert, gesteuert und verantwortet werden.
Ansatz für die KI-Sicherheitsprüfung *AI security testing approach*	Eine strukturierte Methodik zur Planung und Durchführung von Sicherheitstests für ein KI-System, die den Umfang, die Techniken und die Erfolgskriterien abdeckt.
KI-System *AI system*	Ein komplettes System, das auf einem oder mehreren KI-Modellen basiert und Datenpipelines, Schnittstellen sowie die zugehörige Infrastruktur umfasst.
Anbieter eines KI-Systems *AI system provider*	Die Organisation, die ein KI-System entwickelt oder zusammenstellt und es für den Einsatz oder die Nutzung bereitstellt.
Anonymisierung *anonymization*	Die irreversible Verarbeitung personenbezogener Daten, sodass die betroffene Person nicht mehr identifiziert werden kann, auch nicht indirekt.
API (Application Programming Interface) *API (application programming interface)*	Eine definierte Sammlung von Protokollen und Tools, die es verschiedenen Softwareprogrammen oder Komponenten ermöglicht, miteinander zu kommunizieren und Funktionalität auszutauschen.

Begriff	Definition
Audit *audit*	Eine unabhängige Überprüfung eines Systems, eines Prozesses oder einer Organisation, um die Einhaltung von Standards, Richtlinien oder Vorschriften festzustellen.
Augmentierungsdaten *augmentation data*	Externe Daten, die zu einem bestimmten Zeitpunkt im Lebenszyklus eines Modells in dessen Kontext eingefügt werden, um dessen Ergebnisse zu verbessern oder zu erweitern, wie beispielsweise Dokumente, die für Retrieval-Augmented Generation (RAG) abgerufen werden.
automatisierte Aufsicht *automated oversight*	Überwachung oder Steuerung eines Systems, die nicht durch Menschen, sondern durch automatisierte Tools erfolgt.
Automatisierung *automation*	Der Einsatz von Technologie zur Ausführung von Aufgaben oder Prozessen mit reduziertem menschlichem Eingriff.
Bias *bias*	Ein systematischer Verzerrungseffekt bei Daten, Entscheidungen oder Ergebnissen, der häufig unbeabsichtigt durch Datenauswahl oder Designentscheidungen entsteht. Je nach Fall kann ein Bias zu einer ungerechten oder ungenauen Behandlung bestimmter Gruppen oder Fälle führen.
Auswirkungsradius *blast radius*	Das Ausmaß an Schaden, Kompromittierung oder Offenlegung, das entstehen würde, wenn eine bestimmte Ressource, Komponente oder Anmeldedaten erfolgreich angegriffen würde oder ausfallen würde.
Chatbot *chatbot*	Eine Softwareanwendung, die über Text- oder Sprachkonversation mit Nutzern interagiert und dabei die Automatisierung von Antworten ermöglicht, die andernfalls von einem Menschen erbracht werden müssten.
Chief AI Officer (CAIO) *Chief AI Officer (CAIO)*	Die Rolle, die für die gesamte KI-Strategie, die KI-Governance und die Aufsicht über KI-Risiken einer Organisation verantwortlich ist.
Chief Information Security Officer (CISO) *Chief Information Security Officer (CISO)*	Die Führungskraft, die für die gesamte Informations- und Cybersicherheitsstrategie sowie das entsprechende Programm einer Organisation verantwortlich ist.
Code *code*	Die in einer Programmiersprache verfasste Anweisungsfolge, die die Logik einer Software umsetzt.
Compliance *compliance*	Einhaltung der geltenden Gesetze, Vorschriften, Standards und internen Richtlinien, die die Aktivitäten einer Organisation regeln.

Begriff	Definition
Vertraulichkeit *confidentiality*	Die Eigenschaft, dass Informationen nur denjenigen zugänglich sind, die dazu berechtigt sind. Eines der drei Kernziele der Informationssicherheit neben Integrität und Verfügbarkeit.
Konfigurationsleck *configuration leak*	Die unbeabsichtigte Offenlegung von Konfigurationsdetails eines Systems, wie z. B. Einstellungen oder Anmeldedaten, gegenüber unbefugten Personen.
Urheberrecht *copyright*	Eine Form des rechtlichen Schutzes, der den Urhebern von Originalwerken gewährt wird und ihnen ausschließliche Rechte an der Vervielfältigung, Verbreitung und Nutzung dieser Werke einräumt.
Datenleck *data leak*	Die unbefugte oder unbeabsichtigte Offenlegung von Daten gegenüber Personen, die keinen Zugriff darauf haben sollten.
Datenschutz *data protection*	Im Zusammenhang mit personenbezogenen Daten sind dies Maßnahmen und Vorgehensweisen, die darauf abzielen, personenbezogene Daten vor unbefugtem Zugriff, unbefugter Nutzung oder unbefugter Weitergabe zu schützen.
Datenschutz-Folgenabschätzung (DSFA) *data protection impact assessment (DPIA)*	Ein strukturierter Prozess, der nach dem Datenschutzrecht vorgeschrieben ist, um Risiken für die Privatsphäre zu identifizieren und zu mindern, bevor eine Art der Datenverarbeitung durchgeführt wird, die voraussichtlich ein hohes Risiko für betroffene Personen mit sich bringt.
Datensicherheit *data security*	Der Schutz von Daten vor unbefugtem Zugriff, Veränderung oder Verlust während ihres gesamten Lebenszyklus.
betroffene Person *data subject*	Eine identifizierte oder identifizierbare lebende Person, auf die sich personenbezogene Daten beziehen.
Denial-of-Service (DoS) *denial-of-service (DoS)*	Ein Angriff, der darauf abzielt, ein System oder einen Dienst für berechtigte Nutzer unzugänglich zu machen, typischerweise durch Überlastung mit Datenverkehr oder Anfragen.
Denial-of-Wallet (DoW) *denial-of-wallet (DoW)*	Ein Angriff, der die Betriebskosten eines Ziels absichtlich in die Höhe treibt, beispielsweise durch das Auslösen übermäßig ressourcenintensiver Anfragen bei einem nutzungsabhängigen Dienst, wodurch eher ein finanzieller Schaden als ein Verfügbarkeitsverlust entsteht.
Betreiber *deployer*	Eine Person oder Organisation, die ein System in den operativen Einsatz bringt, im Unterschied zu der Partei, die es entwickelt hat.

Begriff	Definition
Entwicklungsphase *development-time*	Die Phase im Lebenszyklus eines Systems, in der es entworfen, erstellt und trainiert bzw. getestet wird, im Gegensatz zu der Phase, in der es bereitgestellt wird und läuft.
• **Development-Time Security Threats (Sicherheitsbedrohungen in der Entwicklungsphase)** *development-time security threats*	Sicherheitsrisiken, die während der Konzeption, der Erstellung oder des Trainings eines Systems entstehen können, wie beispielsweise kompromittierter Code, Komponenten mit Hintertüren oder beschädigte Daten.
• **Data Poisoning during Development-Time (Datenvergiftung in der Entwicklungsphase)** *data poisoning development-time*	Die absichtliche Verfälschung oder Manipulation der Daten, die zum Trainieren eines Modells für maschinelles Lernen verwendet werden, damit das resultierende Modell fehlerhaftes oder von Angreifern gesteuertes Verhalten erlernt.
• **Model Poisoning during Development-Time (Modellvergiftung in der Entwicklungsphase)** *model poisoning development-time*	Direkte Manipulation der Parameter, der Modellarchitektur oder des Trainingsprozesses eines Modells, um verstecktes oder böswilliges Verhalten einzubetten.
• **Sensitive Data Leak during Development-Time (Leck sensibler Daten in der Entwicklungsphase)** *sensitive data leak development-time*	Die unbeabsichtigte Offenlegung sensibler Daten, wie beispielsweise Quelldaten, Anmeldedaten oder Konfigurationsgeheimnisse, während des Entwicklungsprozesses eines Systems.
Embedding *embedding*	Eine numerische Vektordarstellung von Daten, wie beispielsweise Text oder Bilder, die deren semantische Bedeutung erfasst, sodass Modelle des maschinellen Lernens diese verarbeiten, vergleichen oder bei einer Suche nutzen können.
Verschlüsselung *encryption*	Die Umwandlung von Daten in eine unlesbare Form, sodass diese nur von Parteien gelesen werden können, die über den richtigen Entschlüsselungsschlüssel verfügen; dient dem Schutz der Vertraulichkeit von Daten.
Ethikgrundsätze *ethical principles*	Eine Reihe von Werten wie Fairness, Ehrlichkeit und Verantwortung, die als Leitlinien dafür dienen, was in einem bestimmten Bereich oder bei einer bestimmten Tätigkeit als richtiges oder akzeptables Verhalten gilt.

Begriff	Definition
Europäische Union (EU) *European Union (EU)*	Ein politischer und wirtschaftlicher Zusammenschluss europäischer Mitgliedstaaten, der Gesetze und Verordnungen wie die DSGVO und die KI-Verordnung (KI-VO) verabschiedet, die für seine Mitglieder verbindlich sind.
G.U.A.R.D. (Steuern, Verstehen, Anpassen, Reduzieren, Nachweisen) *G.U.A.R.D. (Govern, Understand, Adapt, Reduce, Demonstrate)*	Das fünfstufige Rahmenwerk der OWASP AI Exchange (Govern/Steuern, Understand/Verstehen, Adapt/Anpassen, Reduce/Reduzieren, Demonstrate/Nachweisen) zur Strukturierung des Ansatzes einer Organisation beim Management von KI-Sicherheitsrisiken.
Datenschutz-Grundverordnung (DSGVO) *General Data Protection Regulation (GDPR)*	Die Verordnung der Europäischen Union, die Regeln für den Datenschutz und die Verarbeitung personenbezogener Daten festlegt.
generative KI *generative AI*	KI-Systeme, die neue Inhalte wie Texte, Bilder oder Code erzeugen, in der Regel durch das Erlernen von Mustern aus umfangreichen Trainingsdatensätzen.
Übernahme *hijacking*	Ein Angriff, bei dem ein Angreifer die unbefugte Kontrolle über ein System, eine Sitzung, eine Ressource oder einen Kommunikationskanal übernimmt und diese für seine eigenen Zwecke umleitet.
menschliche Aufsicht *human oversight*	Die Einbindung von Menschen in die Überwachung, Überprüfung oder das Eingreifen in die Entscheidungen und Handlungen eines automatisierten oder autonomen Systems.
Inferenzzeit *inference-time*	Die Phase im Lebenszyklus eines Modells des maschinellen Lernens, in der es aktiv dazu genutzt wird, aus neuen Eingaben Ergebnisse zu generieren – im Gegensatz zur Trainingsphase.
Informationssicherheits-Managementsystem (ISMS) *information security management system (ISMS)*	Ein systematisches Rahmenwerk aus Richtlinien, Prozessen und Sicherheitskontrollen gemäß der Norm ISO/IEC 27001 zum Management der Informationssicherheitsrisiken einer Organisation.
Eingabedaten *input data*	Daten, die einem System oder Prozess zugeführt werden, um eine Ausgabe zu erzeugen. Im Zusammenhang mit KI sind dies die Daten, die als Eingabe für das KI-Modell verwendet werden.

Begriff	Definition
Input Security Threats (Sicherheitsbedrohungen durch Eingabedaten) *input security threats*	Sicherheitsrisiken, die sich aus Daten oder Inhalten ergeben, die einem System als Eingabe bereitgestellt werden, wie beispielsweise böswillige oder manipulierte Inhalte, die darauf abzielen, dessen Verhalten zu verändern.
• **Evasion (Umgehung)** *evasion*	Ein Angriff, bei dem böswillige Eingaben in Form von Daten – im Gegensatz zu Befehlen – erstellt werden, die speziell darauf ausgelegt sind, ein Erkennungs- oder Klassifizierungssystem zu umgehen, sodass dieses ein falsches Ergebnis liefert. Ein Beispiel hierfür ist ein subtil verändertes Verkehrszeichen, das von einem Bildverarbeitungsmodell falsch interpretiert wird.
• **Model Exfiltration (Modellausleitung)** *model exfiltration*	Das unbefugte Extrahieren einer Kopie oder eines funktionalen Äquivalents eines proprietären Modells, häufig durch wiederholte Abfragen.
• **Prompt-Injection (Prompt-Injektion)** *prompt injection*	Ein Angriff, bei dem die an ein generatives KI-Modell gerichteten Anweisungen durch den Prompt oder eingefügte Inhalte manipuliert werden, um das beabsichtigte Verhalten zu überschreiben.
• **Resource Exhaustion (Ressourcenerschöpfung)** *resource exhaustion*	Ein Angriff oder ein Zustand, bei dem Rechenleistung, Arbeitsspeicher oder Speicherplatz in einem solchen Ausmaß beansprucht werden, dass der Dienst für berechtigte Nutzer beeinträchtigt oder verweigert wird oder erhebliche finanzielle Auswirkungen entstehen.
• **Sensitive Data Disclosure through Use (Offenlegung sensibler Daten durch Nutzung)** *sensitive data disclosure through use*	Offenlegung sensibler Trainingsdaten oder anderer Daten durch normale oder missbräuchliche Nutzung des Systems.
Integritätsprüfung *integrity verification*	Überprüfung, ob Daten, Code oder andere digitale Artefakte nicht verändert oder manipuliert wurden, in der Regel unter Verwendung von Techniken wie Prüfsummen, Hash-Werten oder digitalen Signaturen.
geistiges Eigentum *intellectual property (IP)*	Rechtlich geschützte geistige Schöpfungen wie Geschäftsgeheimnisse, Patente, urheberrechtlich geschützte Werke und Marken, an denen ihr Schöpfer oder Eigentümer ausschließliche Rechte besitzt.
ISO/IEC 42001 *ISO/IEC 42001*	Eine internationale Norm, die Anforderungen an ein KI-Managementsystem festlegt, um KI verantwortungsbewusst und sicher zu steuern.
Large Language Model (LLM) *large language model (LLM)*	Ein auf der Grundlage riesiger Textmengen trainiertes Modell des maschinellen Lernens, das natürliche Sprache interpretieren und generieren kann.

Begriff	Definition
Leckage *leakage*	Die unbefugte oder unbeabsichtigte Offenlegung von Informationen, die vertraulich sind oder des Schutzes bedürfen.
maschinelles Lernen *machine learning*	Ein Teilgebiet der KI, bei dem Systeme Muster aus Daten lernen, anstatt explizit mit Regeln programmiert zu werden.
Membership Inference *membership inference*	Ein Angriff, bei dem ermittelt wird, ob ein bestimmter Datensatz Teil der Trainingsdaten eines Modells war, wodurch möglicherweise sensible Informationen über Personen offengelegt werden.
Modellarchitektur *model architecture*	Der strukturelle Aufbau eines Modells, einschließlich seiner Komponenten und deren Verknüpfungen.
Modellinversion *model inversion*	Ein Angriff, bei dem sensible Trainingsdaten durch die Analyse der Modellausgaben rekonstruiert oder approximiert werden.
Modellleck *model leak*	Die unbefugte Offenlegung oder der Diebstahl der Parameter, Gewichte oder der Modellarchitektur eines Modells.
quelloffen *open source*	Software oder andere technische Artefakte, deren zugrunde liegender Quellcode oder deren Komponenten öffentlich zugänglich gemacht werden, sodass jeder sie nutzen, einsehen, ändern und weiterverbreiten kann.
Ausgabedaten *output data*	Das Ergebnis, das von einem System oder Prozess als Reaktion auf eine bestimmte Eingabe erzeugt wird. Im Zusammenhang mit KI handelt es sich dabei in der Regel um die Ausgabe eines KI-Modells.
Ausgabe-Sanitizing *output sanitization*	Das Filtern oder Bereinigen der Ausgabe eines Systems, bevor diese weitergeleitet oder angezeigt wird, um schädliche, sensible oder böswillige Inhalte zu entfernen.
Ausgabevalidierung *output validation*	Überprüfung der Ausgabe eines Systems anhand definierter Kriterien, um sicherzustellen, dass sie korrekt, sicher und für den vorgesehenen Verwendungszweck geeignet ist, bevor darauf reagiert wird oder sie an nachgelagerte Systeme weitergeleitet wird.
personenbezogene Daten *personal data*	Jede Information, die sich auf eine identifizierte oder identifizierbare lebende Person bezieht.

Begriff	Definition
prädiktive KI *predictive AI*	KI-Systeme, die darauf ausgelegt sind, Ergebnisse vorherzusagen oder Daten auf der Grundlage von Mustern zu klassifizieren, die aus historischen Daten gelernt wurden, im Gegensatz zur Generierung neuer Inhalte.
Privatsphäre *privacy*	Das Recht einer Person, zu bestimmen, wie ihre personenbezogenen Daten erhoben, genutzt und weitergegeben werden.
Datenschutz durch Voreinstellung *privacy by default*	Der Grundsatz, dass die Standardeinstellungen eines Systems die Option mit dem höchsten Schutz der Privatsphäre bieten sollten, ohne dass von der betroffenen Person zusätzliche Maßnahmen erforderlich sind; ein begleitendes Konzept zu „Datenschutz durch Technikgestaltung“.
Datenschutz durch Technikgestaltung *privacy by design*	Die Einbindung von Maßnahmen zum Schutz der Privatsphäre in die Architektur und die Prozesse eines Systems bereits in den frühesten Entwicklungsphasen, anstatt diese erst nachträglich hinzuzufügen.
Privacy Impact Assessments (PIA) *privacy impact assessments (PIA)*	Eine Bewertung, mit der die Risiken für die Privatsphäre ermittelt und angegangen werden, die ein Projekt, ein System oder ein Prozess für Einzelpersonen darstellen kann.
Datenschutzvorfall *privacy incident*	Ein Vorfall, bei dem personenbezogene Daten in einer Weise verarbeitet, abgerufen oder offengelegt werden, die gegen Bestimmungen zum Schutz der Privatsphäre verstößt.
Verarbeitung (von personenbezogenen Daten) *processing (of personal data)*	Jeder Vorgang, der an personenbezogenen Daten durchgeführt wird, wie beispielsweise die Erhebung, Speicherung, Nutzung oder Löschung.
Prompt *prompt*	Der Eingabetext oder die Anweisungen, die einem generativen KI-Modell gegeben werden, um eine Antwort zu erhalten.
Pseudonymisierung *pseudonymization*	Das Ersetzen identifizierender Informationen in Daten durch künstliche Identifikatoren, sodass zusätzliche Informationen, die separat gespeichert werden, erforderlich sind, um die Person wieder zu identifizieren.
öffentlich zugänglich *publicly accessible*	Daten, Systeme oder Ressourcen, auf die jeder ohne Authentifizierung oder besondere Berechtigung zugreifen kann.
vorgefertigtes Modell *ready-made model*	Ein vortrainiertes Modell, das von einem Dritten bezogen und unverändert oder mit Anpassungen verwendet wird, anstatt von Grund auf neu trainiert zu werden.

Begriff	Definition
Restrisiko *residual risk*	Das Risiko, das nach der Anwendung von Sicherheitskontrollen und Risikominderungen verbleibt.
verantwortungsvolle KI *responsible AI*	Ein Ansatz zur Entwicklung und zum Betrieb von KI-Systemen, der neben der Leistungsfähigkeit auch Ethik, Fairness, Safety und Sicherheit berücksichtigt.
Retrieval-Augmented Generation (RAG) *retrieval-augmented generation (RAG)*	Eine Technik, bei der ein generatives KI-Modell zur Inferenzzeit relevante externe Dokumente oder Daten abruft, um seine Antworten zu untermauern und anzureichern.
Risiko *risk*	Die Möglichkeit, dass eine Bedrohung eine Sicherheitslücke ausnutzt und Schaden verursacht; dies wird in der Regel anhand der Wahrscheinlichkeit und der Auswirkungen bewertet.
Risikoanalyse *risk analysis*	Der Prozess der Identifizierung und Bewertung von Risiken, um deren Eintrittswahrscheinlichkeit und potenzielle Auswirkungen zu ermitteln.
Risikostufen für KI-Systeme gemäß KI-VO *risk levels for AI systems based on AI Act*	Die Klassifizierung von KI-Systemen in verschiedene Risikostufen gemäß der KI-VO, die bestimmt, welche rechtlichen Verpflichtungen für ein bestimmtes System gelten.
• **minimales oder kein Risiko** *minimal or no risk*	Die Stufe der KI-VO für KI-Systeme, von denen ausgegangen wird, dass sie ein vernachlässigbares Risiko für Gesundheit, Sicherheit oder Grundrechte darstellen und daher kaum oder gar nicht reguliert werden.
• **begrenztes Risiko** *limited risk*	Die Stufe der KI-VO für KI-Systeme, die ein gewisses Risiko darstellen und in der Regel spezifische Pflichten der Transparenz erfordern, wie beispielsweise die Offenlegung, dass Inhalte KI-generiert sind.
• **hohes Risiko** *high risk*	Die Stufe der KI-VO für KI-Systeme, die erhebliche Auswirkungen auf Gesundheit, Sicherheit oder Grundrechte haben können und strengen Verpflichtungen vor und nach der Inbetriebnahme unterliegen.
• **unannehmbares Risiko** *unacceptable risk*	Die Stufe der KI-VO für Praktiken im KI-Bereich, von denen angenommen wird, dass sie ein derart schwerwiegendes Risiko darstellen, dass sie gänzlich verboten sind.
Risikomanagement *risk management*	Der fortlaufende Prozess der Identifizierung, Bewertung, Behandlung und Überwachung von Risiken, um diese auf einem akzeptablen Niveau zu halten.

Begriff	Definition
Risikominderung *risk mitigation*	Die Auswahl und Umsetzung von Maßnahmen zur Verringerung der Eintrittswahrscheinlichkeit oder der Auswirkungen eines identifizierten Risikos.
Risikoeinstufung *risk rating*	Eine Bewertung oder Klassifizierung, die einem Risiko auf der Grundlage seiner eingeschätzten Eintrittswahrscheinlichkeit und seiner Auswirkungen zugewiesen wird.
Runtime *runtime*	Die Phase im Lebenszyklus eines Systems, in der es bereitgestellt ist und aktiv betrieben wird, wobei es reale Eingaben in einer Live-Umgebung verarbeitet.
Runtime Security Threats (Sicherheitsbedrohungen während der Runtime) *runtime security threats*	Sicherheitsrisiken, die auftreten, während ein System in einer Live-Produktionsumgebung betrieben wird, im Gegensatz zur Entwicklungsphase.
• **Augmentation Data Manipulation (Manipulation von Augmentierungsdaten)** *augmentation data manipulation*	Manipulation der externen Daten, die ein System während der Runtime abruft oder einspeist (beispielsweise bei RAG-Quellen), um dessen Ausgabe zu beeinflussen.
• **Direct Augmentation Data Leak (direktes Leck von Augmentierungsdaten)** *direct augmentation data leak*	Die Offenlegung von Augmentierungsdaten durch unbefugten direkten Zugriff während der Übertragung oder im Ruhezustand.
• **Direct Runtime Model Leak (direktes Modellleck während der Runtime)** *direct runtime model leak*	Die unbefugte Offenlegung der Parameter oder der Modellarchitektur eines Modells durch unbefugten direkten Zugriff während der Übertragung oder im Ruhezustand.
• **Direct Runtime Model Poisoning (direkte Modellvergiftung während der Runtime)** *direct runtime model poisoning*	Die Manipulation des Verhaltens eines bereitgestellten Modells durch einen Angreifer mittels unbefugten direkten Zugriffs während der Runtime, während der Übertragung oder im Ruhezustand.
• **Input Data Leak (Leck von Eingabedaten)** *input data leak*	Die unbefugte Offenlegung von Daten, die einem System als Eingabe bereitgestellt wurden.

Begriff	Definition
• **Output containing Conventional Injection (Ausgabe mit klassischer Injection)** *output containing conventional injection*	Ausgabe eines Systems, die selbst schädliche Inhalte wie Skript- oder Abfragefragmente enthält und bei der Weiterleitung an ein nachgelagertes System einen konventionellen Injektionsangriff wie SQL-Injection oder Cross-Site-Scripting auslöst.
Safety eines KI-Systems *safety of an AI system*	Die Eigenschaft, dass ein KI-System durch seinen Betrieb keinen inakzeptablen Schaden an Menschen, Eigentum oder der Umwelt verursacht.
Sicherheitsverstoß *security breach*	Ein Vorfall, bei dem Sicherheitskontrollen umgangen werden oder versagen, was zu unbefugtem Zugriff, Offenlegung oder Schaden führt.
Sicherheitskontrollen *security controls*	Die technischen, verfahrensbezogenen oder organisatorischen Maßnahmen, die zur Minderung von Sicherheitsrisiken getroffen werden.
• **Kontrollen zur Begrenzung sensibler Daten** *controls for sensitive data limitation*	Sicherheitskontrollen, die speziell darauf abzielen, die Erfassung, Offenlegung oder Speicherung sensibler Daten einzuschränken.
• **Kontrollen zur Begrenzung der Auswirkungen unerwünschten Verhaltens** *controls to limit the effect of unwanted behavior*	Sicherheitskontrollen, die darauf ausgelegt sind, die Auswirkungen unbeabsichtigter oder schädlicher Verhaltensweisen eines Systems einzudämmen oder zu verringern, sobald diese auftreten.
• **Governance-Kontrollen** *governance controls*	Kontrollen auf Organisationsebene, wie Richtlinien, Strukturen der Aufsicht und Mechanismen der Verantwortung, die Risiken auf strategischer Ebene steuern.
(Sicherheits-)Angriff *(security) attack*	Eine vorsätzliche Handlung eines Angreifers mit der Absicht, die Vertraulichkeit, Integrität oder Verfügbarkeit eines Systems oder seiner Assets zu beeinträchtigen.
Sicherheitsvorfall *security incident*	Ein Ereignis, das auf einen möglichen Verstoß gegen Sicherheitsrichtlinien, ein Versagen von Kontrollmaßnahmen oder die Kompromittierung einer Ressource hindeutet.
Sicherheitsrisiken *security risks*	Das Potenzial von Sicherheitsbedrohungen, Sicherheitslücken auszunutzen und Schaden anzurichten.
Sicherheitstests *security testing*	Der Prozess der Bewertung eines Systems, um Sicherheitsschwächen zu identifizieren, bevor diese ausgenutzt werden können.

Begriff	Definition
Sicherheitsbedrohungen *security threats*	Die vollständige Auflistung aller Möglichkeiten, wie die Vertraulichkeit, Integrität oder Verfügbarkeit eines Systems angegriffen werden kann. Im Zusammenhang mit KI-Systemen umfasst dies Bedrohungen für Eingabedaten, Bedrohungen während der Entwicklungsphase (durch Eingaben) sowie herkömmliche Bedrohungen während der Runtime.
sensible Daten *sensitive data*	Daten, die aufgrund des potenziellen Schadens bei einer Offenlegung besonderen Schutz erfordern, wie beispielsweise personenbezogene, vertrauliche oder geheime Informationen.
Service Provider *service provider*	Ein Dritter, der eine Dienstleistung wie Hosting, Infrastruktur oder Software bereitstellt, auf die eine Organisation oder ein System angewiesen ist.
Quellcode *source code*	Die in einer Programmiersprache verfassten, für Menschen lesbaren Anweisungen, aus denen eine Software besteht.
Quellcode-Leck *source code leak*	Die unbefugte Offenlegung des Quellcodes eines Systems gegenüber Unbefugten.
Sponge-Angriff *sponge attack*	Ein Angriff, bei dem einem Modell Eingaben zugeführt werden, die darauf ausgelegt sind, dessen Rechenzeit oder Ressourcenverbrauch zu maximieren, wodurch dessen Leistung oder Verfügbarkeit beeinträchtigt wird.
Stakeholder *stakeholder*	Jede Person oder Organisation, die ein Interesse an einem Projekt, einem System oder einer Entscheidung hat oder davon betroffen ist.
Start-up-Unternehmen *startup*	Ein Unternehmen in der Frühphase, das in der Regel gegründet wurde, um ein neues Produkt, eine neue Technologie oder ein neues Geschäftsmodell zu entwickeln und zu skalieren.
Structured Query Language (SQL) *structured query language (SQL)*	Eine Standardsprache, die zum Erstellen, Abfragen und Verwalten von Daten in relationalen Datenbanken verwendet wird.
Manipulation *tampering*	Unbefugte Änderung von Daten, Code oder einem System, um dessen Verhalten oder Integrität zu verändern.
Dritter *third party*	Jede externe Organisation oder Person, die nicht der primäre Anbieter oder Nutzer ist, aber dennoch an einem System beteiligt ist oder dieses beeinflusst, beispielsweise über dessen Lieferkette.
Bedrohungsmodellierung *threat modeling*	Ein strukturierter Prozess zur Identifizierung potenzieller Bedrohungen, Sicherheitslücken und Angriffswege auf

Begriff	Definition
	ein System, der zur Priorisierung von Sicherheitsmaßnahmen dient.
Token *token*	Die Grundeinheit eines Textes, wie beispielsweise ein Wort, ein Teil eines Wortes oder ein Zeichen, in die ein Text zur Verarbeitung durch ein Sprachmodell zerlegt wird. Im Zusammenhang mit KI-Cloud-Diensten sind Token zudem die Abrechnungseinheit.
Tokenisierung *tokenization*	Der Vorgang, bei dem Text zur Verarbeitung durch ein Sprachmodell in kleinere Einheiten aufgeteilt wird oder bei dem sensible Daten durch nicht sensible Platzhalterwerte ersetzt werden, um die Originaldaten zu schützen.
Tool *tool*	Im Kontext der agentischen KI eine externe Funktion, eine API (Application Programming Interface) oder eine Fähigkeit, die ein KI-Agent aufrufen kann, um über die Textgenerierung hinausgehende Aktionen auszuführen.
Trainingsdaten *training data*	Die Daten, die verwendet werden, um ein Modell des maschinellen Lernens darauf zu trainieren, Muster zu erkennen oder seine Aufgabe auszuführen.
vertrauenswürdige KI *trustworthy AI*	Eine KI, die so konzipiert und betrieben wird, dass sie rechtmäßig, ethisch und technisch robust ist, einschließlich Sicherheit und Safety.
unerwünschtes Verhalten *unwanted behavior*	Jede Ausgabe, Entscheidung oder Aktion eines Systems, die von seinem beabsichtigten, sicheren oder erwarteten Verhalten abweicht.
Sicherheitslücke *vulnerability*	Eine Schwachstelle in einem System, einem Prozess oder einer Kontrollmaßnahme, die von einer Bedrohung ausgenutzt werden könnte, um Schaden anzurichten.

Certified for what's next

Contact EXIN

www.exin.com