

Musterprüfung

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.

Inhalt

Einführung	4
Musterprüfung	5
Antwortschlüssel	21
Beurteilung	57

Einführung

Dies ist die EXIN AI Security Professional based on the OWASP AI Exchange (AISP.DE) Musterprüfung. Es gilt die Prüfungsordnung von EXIN.

Die Musterprüfung besteht aus 40 Multiple-Choice-Fragen. Zu jeder Multiple-Choice-Frage werden mehrere Antwortmöglichkeiten angeboten. Es gibt jeweils eine richtige Antwort.

Sie können maximal 40 Punkte erreichen. Jede richtige Antwort zählt 1 Punkt. Um die Prüfung zu bestehen, müssen Sie mindestens 26 Punkte erzielen.

Die Bearbeitungszeit beträgt 90 Minuten.

Viel Erfolg!

Musterprüfung

1 / 40
Ein IT-Sicherheitsmanager ist bei einem Logistikunternehmen tätig. Das Unternehmen hat kürzlich beschlossen, KI-Tools zur Optimierung seiner Lieferrouten einzusetzen. Der IT-Sicherheitsmanager ist für die Organisation der KI-Sicherheit verantwortlich. Die ersten beiden Schritte des G.U.A.R.D.-Frameworks sind bereits abgeschlossen:

- Der erste Schritt Govern (Steuern) wurde durch die Erstellung einer KI-Sicherheitsrichtlinie umgesetzt.
- In der zweiten Stufe Understand (Verstehen) wurden alle KI-Assets und die damit verbundenen Risiken erfasst.

Was ist der **nächste** Schritt, den der IT-Sicherheitsmanager unternehmen sollte?

A) Anpassung der Sicherheitspraktiken um KI-spezifische Bedrohungsmodellierung, Tests und Kontrollen der Lieferkette
B) Analyse der identifizierten KI-Risiken und Festlegung einer Strategie zur Risikopriorisierung entsprechend dem potenziellen Schadensausmaß
C) Anwendung von Sicherheitskontrollen auf alle KI-Systeme zur Gewährleistung einer verantwortungsvollen KI-Sicherheit entlang des gesamten KI-Lebenszyklus
D) Auswertung der im Schritt Govern (Steuern) erstellten KI-Sicherheitsrichtlinie, um die erfassten KI-Assets und deren Risiken einzubeziehen

2 / 40
Ein Unternehmen überprüft seine KI-Strategie, um sicherzustellen, dass diese einen geschäftlichen Mehrwert schafft und gleichzeitig ethischen sowie operativen Anforderungen gerecht wird. Der KI-Sicherheitsbeauftragte unterscheidet dabei zwischen den beiden zentralen Dimensionen verantwortungsvolle KI und vertrauenswürdige KI.

Was ist der **wesentliche** Unterschied zwischen verantwortungsvoller KI und vertrauenswürdiger KI?

A) - Verantwortungsvolle KI konzentriert sich auf Ethik, Gesellschaft und Governance.
- Vertrauenswürdige KI konzentriert sich auf technische und betriebliche Aspekte, Transparenz und Erklärbarkeit.
B) - Verantwortungsvolle KI konzentriert sich auf die Compliance mit einschlägigen Gesetzen und Rechtsvorschriften.
- Vertrauenswürdige KI konzentriert sich auf Datenschutz und Cybersicherheit, um den Datenschutzgrundsätzen zu entsprechen.
C) - Verantwortungsvolle KI konzentriert sich auf die Systemgenauigkeit und technische Leistungskennzahlen.
- Vertrauenswürdige KI konzentriert sich auf ethische Leitlinien und gesellschaftliche Auswirkungen, um das Vertrauen der Verbraucher zu gewinnen.
D) - Verantwortungsvolle KI konzentriert sich auf die Anbieterauswahl, Risiken und das Kostenmanagement.
- Vertrauenswürdige KI konzentriert sich auf eine sichere und zuverlässige Schulung der Nutzer sowie die Einführung von KI-Systemen.

3 / 40
Ein Cybersicherheitsteam verfügt bereits über ein Framework zur Erkennung von Bedrohungen. Das Team muss nun auch ein neues Large Language Model (LLM) schützen, das im Kundensupport für einen Chatbot eingesetzt wird.

Das Team möchte dieselben Tools verwenden, die es auch für Webanwendungen einsetzt:
- Signaturbasierte Erkennungsregeln
- Feste Kontrollen zur Eingabevalidierung

Warum reicht dieser Ansatz **nicht** aus, um das LLM zu sichern?

A) Weil KI-Systeme andere Angriffsflächen aufweisen und spezifische Kontrollen für das Verhalten des LLM sowie die Datenpipeline erfordern
B) Weil das Framework zur Erkennung von Bedrohungen keine Verfahren zur Reaktion auf Sicherheitsvorfälle umfasst und daher für LLMs ungeeignet ist
C) Weil LLMs Ausgaben generieren, die zu variabel sind, um von regelbasierten Systemen bewertet werden zu können, und daher von einem Menschen überprüft werden müssen
D) Weil LLMs Anfragen schneller verarbeiten als Webanwendungen, sodass signaturbasierte Regeln nicht mit ihnen Schritt halten können
E) Weil Firewalls für Webanwendungen (Web Application Firewall, WAF) nicht für den Einsatz mit LLMs lizenziert sind und separate Vereinbarungen mit den Anbietern erfordern

4 / 40
Bei der Nutzung von KI verfügt ein Unternehmen über KI-Assets. Für jedes KI-Asset lassen sich spezifische Hauptbedrohungen identifizieren.

Welche Kombination aus einem KI-Asset und seiner Hauptbedrohung ist **korrekt**?

A) - Asset: Von Netzwerk-Firewalls generierte Audit-Protokolle
- Bedrohung: AI Resource Exhaustion (KI-Ressourcenerschöpfung)
B) - Asset: Authentifizierungsdaten von Nutzern
- Bedrohung: Die Ausgabe enthält eine klassische Injection
C) - Asset: In einem Versionskontrollsystem gespeicherter Software-Quellcode
- Bedrohung: Direct Runtime Model Poisoning (direkte Modellvergiftung während der Runtime)
D) - Asset: Trainingsdaten, die zum Aufbau eines Modells verwendet werden
- Bedrohung: Offenlegung sensibler Daten in der Ausgabe

5 / 40
Eine Anwaltskanzlei nutzt zur Dokumentenprüfung und -analyse ein KI-System. Das Sicherheitsteam arbeitet mit einem Risikomanagementansatz. Eine Bestandsaufnahme aller potenziellen Bedrohungen für das KI-System wurde durchgeführt.

Welchen Schritt sollte das Team als **Nächstes** unternehmen?

A) Einen Prozess der Bedrohungsmodellierung anwenden, um die Bedrohungen in eine Reihe konkreter und priorisierter Risiken zu überführen.
B) Jeder identifizierten Bedrohung einen Risikoverantwortlichen zuweisen, um von Anfang an „Protection by Design“ zu gewährleisten.
C) Für alle Bedrohungen Sicherheitskontrollen implementieren, um weitere Risiken entlang des gesamten Lebenszyklus des KI-Systems zu mindern.
D) Die vollständige Liste der Bedrohungen an die relevanten Stakeholder weitergeben, um das Bewusstsein zu schärfen und Transparenz zu schaffen.

6 / 40
Ein KI-Sicherheitsingenieur überprüft das Design der Zugriffskontrolle für ein KI-Multi-Agenten-System. Alle Agenten im System teilen sich ein Dienstkonto. Dieses gemeinsame Dienstkonto hat Lese- und Schreibzugriff auf die Produktionsdatenbank, das interne Dateisystem und das externe API-Gateway (Application Programming Interface).

Die verschiedenen KI-Agenten rufen autonom Tools auf und führen mehrstufige Workflows über interne Systeme hinweg aus. Die Bereitstellung erfolgte über ein gemeinsames Dienstkonto, um die Verwaltung der Anmeldedaten entlang der gesamten KI-Agent-Pipeline zu vereinfachen.

Warum ist dieses Design mit einem gemeinsamen Dienstkonto bei einer agentischen KI **besonders riskant**?

A) Weil KI-Agenten ein höheres Volumen an API-Aufrufen generieren als menschliche Nutzer, sodass es wahrscheinlicher ist, dass Ratenbegrenzungsmaßnahmen ausgelöst werden und das gemeinsam genutzte Dienstkonto gesperrt wird
B) Weil KI-Agenten keine Multi-Faktor-Authentifizierung verwenden, sodass die gemeinsam genutzten Anmeldedaten einen Single Point of Failure ohne weitere Verifizierungsebenen darstellen, den ein Angreifer ausnutzen kann
C) Weil KI-Agenten innerhalb einer einzigen Sitzung autonom über mehrere Schlussfolgerungsschritte hinweg agieren, sodass ein kompromittierter Agent Aktionen über alle zugänglichen Systeme hinweg verketten kann
D) Weil regulatorische Rahmenbedingungen für KI-Systeme vorschreiben, dass jeder KI-Agent mit eindeutigen Anmeldedaten ausgestattet sein muss, sodass gemeinsam genutzte Dienstkonten einen direkten Compliance-Verstoß darstellen
E) Weil die gemeinsame Nutzung von Dienstkonten durch mehrere KI-Agenten während der vorgeschriebenen Rotation der Anmeldedaten und bei Aktivitäten im Zugriffsmanagement zu einer zu hohen betrieblichen Komplexität führt

7 / 40
Ein Ingenieur für maschinelles Lernen bei einem Finanzinstitut untersucht zwei separate Vorfälle im Zusammenhang mit dem Betrugserkennungsmodell des Instituts.

- **Vorfall 1**: Ein externer Angreifer, der keinen Zugriff auf die Modellarchitektur, die Trainingsdaten oder die Parameter des Modells hat, übermittelt Tausende von manipulierten Transaktionen. Diese werden iterativ verfeinert, je nachdem, ob der Betrugsalarm ausgelöst wurde oder nicht.
- **Vorfall 2**: Ein Forscher verwendet ein lokal trainiertes Ersatzmodell mit ähnlicher Modellarchitektur, um adversariale Transaktionen zu generieren. Diese umgehen das Produktionsmodell erfolgreich, ohne dass zuvor eine Interaktion stattgefunden hat.

Um welche zwei Arten von Evasion Attacks (Umgehungsangriffe) handelt es sich hier?

A) - Vorfall 1: Evasion After Poisoning (Umgehung nach Vergiftung)
- Vorfall 2: Perfect-Knowledge Evasion (Umgehungsangriff mit vollständigem Wissen)
B) - Vorfall 1: Partial-Knowledge Evasion (Umgehungsangriff mit Teilwissen)
- Vorfall 2: Transfer Attack (Transferangriff)
C) - Vorfall 1: Perfect-Knowledge Evasion
- Vorfall 2: Zero-Knowledge Evasion (Umgehungsangriff mit Nullwissen)
D) - Vorfall 1: Transfer Attack
- Vorfall 2: Partial-Knowledge Evasion
E) - Vorfall 1: Zero-Knowledge Evasion
- Vorfall 2: Transfer Attack

8 / 40
Erick ist Nutzer einer Social-Media-Plattform, die Kommentare mithilfe eines KI-Systems automatisch moderiert. Er versucht, einen gegen die Richtlinien verstoßenden Kommentar an der automatisierten Moderation der Plattform vorbeizuschleusen, da er die Interaktion mit seinem Beitrag steigern möchte. Erick hat keinerlei Informationen über die Architektur, die Parameter oder die Trainingsdaten des Modells und sieht beim Versuch, seinen Kommentar zu veröffentlichen, lediglich eine binäre Meldung: entweder „blockiert“ oder „veröffentlicht“. Indem er wiederholt kleine Änderungen am Wortlaut seines Kommentars vornimmt, findet Erick schließlich eine Version, die veröffentlicht wird.

Um welche Art von Evasion Attack (Umgehungsangriff) handelt es sich hier?

A) Zero-Knowledge Evasion (Umgehungsangriff mit Nullwissen)
B) Partial-Knowledge Evasion (Umgehungsangriff mit Teilwissen)
C) Perfect-Knowledge Evasion (Umgehungsangriff mit vollständigem Wissen)
D) Evasion After Poisoning (Umgehung nach Vergiftung)
E) Transfer Attack (Transferangriff)

9 / 40
Ein Rechtsdienstleistungsunternehmen entwickelt einen Chatbot, der Fragen von Nutzern beantwortet und Informationen aus internen und externen Quellen wie Websites und Unternehmensdokumenten abrufen kann. Während einer Demo löst das Verhalten des Chatbots eine Untersuchung durch die KI-Sicherheit aus. Sie identifiziert zwei separate Bedrohungen:

- **Bedrohung 1:** Eine Nachricht, die so gestaltet ist, dass sie die Anweisungen und Sicherheitsregeln des Chatbots außer Kraft setzt und dazu führen kann, dass der Chatbot den Befehlen des Testers folgt, anstatt sich an die konfigurierten Richtlinien zu halten.
- **Bedrohung 2:** Eine vom Chatbot abgerufene Anbieter-Website enthält eingebetteten Text mit Anweisungscharakter. Wenn der Chatbot diese Seite aufruft, behandelt er den Inhalt als Befehle und weicht von seinen Richtlinien ab.

Um welche Arten von Bedrohungen handelt es sich?

A) - Bedrohung 1: Evasion Attack (Umgehungsangriff)
- Bedrohung 2: Data Poisoning (Datenvergiftung)

B) - Bedrohung 1: Data Poisoning
- Bedrohung 2: direkte Prompt-Injection

C) - Bedrohung 1: direkte Prompt-Injection
- Bedrohung 2: indirekte Prompt-Injection

D) - Bedrohung 1: indirekte Prompt-Injection
- Bedrohung 2: Evasion Attack

10 / 40
Ein Beratungsunternehmen setzt zur Unterstützung bei alltäglichen Aufgaben ein KI-System ein. Nach einem Gespräch mit einem Kunden fasst der KI-Assistent das Gespräch zusammen. Der KI-Assistent beginnt, vertrauliche interne Links in den Entwurf einer E-Mail einzufügen, die an den Kunden gesendet werden soll. Ein autonomer Anomaliedetektor unterbricht jedoch den Versand und stellt den Entwurf unter Quarantäne.

Der KI-Sicherheitsarchitekt möchte den Zugriff des KI-Besprechungsassistenten vorab auf die Berechtigungen der betreuten Personen beschränken.

Welche Schichten des Prompt-Injection-Schutzes werden hier dargestellt?

A) Automatisierte Aufsicht und benutzerbasiertes Prinzip der geringsten Berechtigungen
B) Just-in-Time-Autorisierung und menschliche Aufsicht
C) Modellabgleich und absichtsbasiertes Prinzip der geringsten Berechtigungen
D) Verarbeitung von Eingaben und Ausgaben bei Prompt-Injection sowie Ratenbegrenzung

11 / 40
Ein Prüfer bewertet die Datenschutzrisiken eines im Gesundheitswesen eingesetzten Large Language Models (LLM) sowie dessen Compliance mit den Datenschutzgesetzen. Zu diesem Zweck testet der Prüfer das Modell mit gezielten Abfragen zu einem bestimmten Patienten und sorgfältig abgestimmten künstlich erzeugten Nichtpatienten. Durch die Beobachtung von Unterschieden in den Antwortmustern und Konfidenzwerten des Modells kann der Prüfer zuverlässig feststellen, ob die Akte dieses Patienten in den Trainingsdaten enthalten war, auch wenn keine tatsächlichen medizinischen Details preisgegeben werden.

Welche Art von Angriff wird hier simuliert?

A) Datenoffenlegung in der Modellausgabe
B) Indirekte Prompt-Injection
C) Membership Inference
D) Model Inversion (Modellinversion)

12 / 40
Was zählt zu den **wesentlichen** Folgen eines erfolgreichen Angriffs auf die Model Exfiltration (Modellausleitung)?

A) Der Angreifer kann den ursprünglichen Trainingsdatensatz in Rohform extrahieren, der zur Erstellung des ursprünglichen Modells verwendet wurde und der möglicherweise sensible Datensätze enthält.
B) Der Angreifer erlangt direkte Kontrolle über das bereitgestellte Modell und kann dessen Gewichte und Parameter ohne vorherige Genehmigung ändern.
C) Der Angreifer erhält eine Kopie des Modells. Diese kann zur Entwicklung von Evasion Attacks (Umgehungsangriffe) genutzt werden, ohne die Abwehrmechanismen des Originals auszulösen.
D) Das ursprüngliche Modell steht legitimen Nutzern aufgrund von Resource Exhaustion (Ressourcenerschöpfung) nicht mehr zur Verfügung, was zu einer Verschlechterung der Dienstqualität und wiederholten Zeitüberschreitungen führt.

13 / 40
Ein in der Cloud gehostetes Large Language Model (LLM) wird Opfer einer Sponge Attack (Sponge-Angriff). Angreifer senden sehr lange oder komplexe Eingaben, die zusätzliche Rechenleistung erfordern. Dies treibt die Kosten in die Höhe und verursacht Verzögerungen oder Ausfälle. Um dieser Situation entgegenzuwirken, möchte der KI-Sicherheitsingenieur zwei Kontrollmaßnahmen einführen:

- **Kontrollmaßnahme 1** validiert und bereinigt Eingaben, um böswillige Inhalte, wie beispielsweise ungewöhnlich große Eingaben, zurückzuweisen oder zu korrigieren.
- **Kontrollmaßnahme 2** begrenzt die pro Modelleingabe verbrauchten Rechenressourcen, um eine Überbeanspruchung zu verhindern.

Um welche beiden Kontrollmaßnahmen handelt es sich?

A) - Kontrollmaßnahme 1: Behandlung anomaler Eingaben
- Kontrollmaßnahme 2: Verschleierung von Konfidenzgrenzen

B) - Kontrollmaßnahme 1: Validierung von Eingaben im Hinblick auf Denial-of-Service (DoS)
- Kontrollmaßnahme 2: Ressourcenbegrenzung

C) - Kontrollmaßnahme 1: Zugriffskontrolle auf das Modell
- Kontrollmaßnahme 2: gegen Evasion Attacks (Umgehungsangriffe) robustes Modell

D) - Kontrollmaßnahme 1: Nutzungsüberwachung
- Kontrollmaßnahme 2: Ratenbegrenzung

14 / 40
Ein KI-Sicherheitsingenieur identifiziert die folgenden Bedrohungen für ein Unternehmen:

- **Adversariale Prompts** werden von Angreifern erstellt, um Fehlklassifizierungen zu verursachen.
- Eine **kompromittierte Erfassungspipeline** führt Datensätze mit Hintertüren in den Korpus ein, bevor das KI-Modell erstellt wird.
- Angreifer manipulieren **Labeling-Workflows** und ändern die Ground-Truth-Annotationen, um das Modellverhalten zu verzerren.
- **Trainingsdaten** werden ohne ordnungsgemäße Validierung aus externen oder Drittquellen bezogen.

Bei welcher Bedrohung besteht laut OWASP AI Exchange **kein** Risiko von Data Poisoning (Datenvergiftung) in der Entwicklungsphase?

A) Adversariale Eingaben
B) Kompromittierte Erfassungspipeline
C) Labeling-Workflows
D) Trainingsdaten

15 / 40
Was ist laut OWASP AI Exchange ein Beispiel für **Direct Development-Time Model Poisoning** (direkte Modellvergiftung in der Entwicklungsphase)?

A) Angreifer erstellen während der Runtime adversariale Eingaben, um das Modell in die Irre zu führen
B) Angreifer manipulieren Gewichte und eine Trainingspipeline, um das Verhalten des Modells zu verändern.
C) Sensible Trainingsdaten werden unbeabsichtigt über die Modellausgaben offengelegt.
D) Ein plötzlicher Anstieg des Datenverkehrs überlastet den KI-Dienst, was zu Verlangsamungen oder Ausfallzeiten führt.

16 / 40
Was ist laut OWASP AI Exchange ein Beispiel für **Supply-Chain Model Poisoning** (Modellvergiftung in der Lieferkette)?

A) Angreifer kompromittieren Daten, Modelle oder Komponenten der Lieferkette vor deren Integration.
B) Angreifer manipulieren Eingaben während der Runtime, um das Modell dazu zu bringen, falsche Ergebnisse zu erzeugen.
C) Das Modell liefert aufgrund unzureichender Daten oder Trainingsqualität falsche Ergebnisse.
D) Das Modell gibt unbeabsichtigt sensible Trainingsdaten über seine Antworten preis.

17 / 40
Ein KI-Entwicklungsteam führt eine Bewertung des Risikos seiner Entwicklungsumgebung durch. Es identifiziert zwei unterschiedliche Bedrohungen:

Bedrohung 1 betrifft den unbefugten Zugriff auf die zum Trainieren und Testen des KI-Modells verwendeten Datensätze.
Bedrohung 2 betrifft den unbefugten Zugriff auf die Modellparameter und -gewichte, die in einem kompromittierten Repository gespeichert sind.

Um welche Arten von Bedrohungen handelt es sich?

A) - Bedrohung 1: Development-Time Data Leak (Datenleck in der Entwicklungsphase)
- Bedrohung 2: Direct Development-Time Model Leak (direktes Modellleck in der Entwicklungsphase)
B) - Bedrohung 1: Direct Augmentation Data Leak (direktes Leck von Augmentierungsdaten)
- Bedrohung 2: Repository Leak (Repository-Leck)
C) - Bedrohung 1: Direct Runtime Data Leak (direktes Datenleck während der Runtime)
- Bedrohung 2: Direct Runtime Model Leak (direktes Modellleck während der Runtime)
D) - Bedrohung 1: Input Data Leak (Leck von Eingabedaten)
- Bedrohung 2: Source Code/Configuration Leak (Quellcode-Leck/Konfigurationsleck)

18 / 40
Ein Sicherheitsteam überprüft die Runtime-Bedrohungslandschaft eines KI-Systems. Das KI-System ist in eine Webanwendung integriert. Das Team identifiziert zwei herkömmliche Sicherheitsbedrohungen:

- **Bedrohung 1:** Ein Angreifer kann Zugriff auf den Speicher des Produktionsservers erlangen und die Gewichte und Parameter des Modells extrahieren.
- **Bedrohung 2:** Ein Angreifer mit Zugriff auf das Modellregister ersetzt das bereitgestellte Modellartefakt durch eine mit einem Trojaner versehene Version.

Welche **KI-spezifischen** Folgen haben diese herkömmlichen Sicherheitsbedrohungen?

A) - Bedrohung 1 führt dazu, dass das System anhand schädlicher Daten automatisch neu trainiert wird.
- Bedrohung 2 zwingt das Modell dazu, gültige Prompts abzulehnen.
B) - Bedrohung 1 verändert die Modellarchitektur während der Runtime.
- Bedrohung 2 komprimiert die Gewichte unbemerkt auf eine geringere Präzision.
C) - Bedrohung 1 deaktiviert Sicherheitsfilter, wodurch das System von Grund auf toleranter wird.
- Bedrohung 2 macht die Ausgaben deterministisch und vorhersehbar.
D) - Bedrohung 1 ermöglicht Modelldiebstahl und potenzielle Rückschlüsse auf Trainingsdaten.
- Bedrohung 2 ermöglicht unentdeckte Manipulationen am Modell.

19 / 40
Ein Finanzdienstleistungsunternehmen setzt ein Modell für maschinelles Lernen (ML-Modell) ein, um betrügerische Transaktionen zu erkennen. Das ML-Modell wird in der Produktionsumgebung des Unternehmens gehostet und verarbeitet kontinuierlich Live-Transaktionsdaten. Es treten zwei Sicherheitsvorfälle auf:

- **Vorfall 1**: Ein Angreifer verschafft sich unbefugt Zugriff auf die Produktionsumgebung des Modells und verändert die Modellparameter so, dass bestimmte betrügerische Transaktionen nicht mehr als solche gekennzeichnet werden. Die Manipulation wird nicht erkannt und der Angreifer nutzt dies zu seinem Vorteil.
- **Vorfall 2**: Ein Angreifer nutzt eine Sicherheitslücke im Produktionssystem und kann so das Modell replizieren und dessen Verhalten für weitere Angriffe analysieren.

Welche Arten von Sicherheitsvorfällen nach OWASP AI Exchange werden hier beschrieben?

A) - Vorfall 1: Data Poisoning (Datenvergiftung)
- Vorfall 2: Direct Runtime Model Poisoning (direkte Modellvergiftung während der Runtime)

B) - Vorfall 1: Data Poisoning
- Vorfall 2: Evasion Attack (Umgehungsangriff)

C) - Vorfall 1: Direct Runtime Model Poisoning
- Vorfall 2: Direct Runtime Model Leak (direktes Modellleck während der Runtime)

D) - Vorfall 1: Direct Runtime Model Leak
- Vorfall 2: Evasion Attack

20 / 40
Eine Einrichtung des Gesundheitswesens setzt einen KI-Assistenten ein, der dem medizinischen Fachpersonal dabei hilft, Patientenakten zusammenzufassen und Empfehlungen auszusprechen. Das System ist mit internen Datenbanken und externen Tools vernetzt und das medizinische Fachpersonal stützt sich im täglichen Arbeitsablauf auf die von ihm generierten Ergebnisse. Es werden zwei Bedrohungen beobachtet:

- **Bedrohung 1**: Der KI-Assistent generiert eine Antwort, die ein verstecktes, bösartiges Skript enthält, das in HTML-Inhalte eingebettet ist. Bei Anzeige der Ausgabe wird das Skript ausgeführt und sendet Session-Daten an den Angreifer.
- **Bedrohung 2**: Ein Angreifer nutzt schwache Zugriffskontrollen im System aus, um sensible Patientendaten abzufangen und abzurufen.

Welche Bedrohungsarten nach OWASP AI Exchange werden hier beschrieben?

A) - Bedrohung 1: Data Poisoning (Datenvergiftung)
- Bedrohung 2: Input Data Leak (Leck von Eingabedaten)

B) - Bedrohung 1: Data Poisoning
- Bedrohung 2: Model Inversion (Modellinversion)

C) - Bedrohung 1: Ausgabe mit klassischer Injection
- Bedrohung 2: Leck von Eingabedaten

D) - Bedrohung 1: Ausgabe mit klassischer Injection
- Bedrohung 2: Model Inversion

21 / 40
Ein Unternehmen nutzt ein RAG-System (Retrieval-Augmented Generation), um Fragen von Mitarbeitenden zu internen Richtlinien zu beantworten. Das KI-System ruft Dokumente aus einer Vektordatenbank ab und fügt diese dem Prompt hinzu, bevor es sie an das KI-Modell übermittelt.

Bei einer Sicherheitsüberprüfung wird festgestellt, dass sich ein Angreifer unbefugten Zugriff auf die Vektordatenbank verschafft hat. Der Angreifer hat die gespeicherten Augmentierungsdokumente eingesehen und ein legitimes Richtliniendokument durch ein anderes ersetzt, das falsche Informationen enthält. Mitarbeitende, die das KI-System zu dieser Richtlinie befragen, erhalten nun falsche Hinweise.

Welche Sicherheitsbedrohungen veranschaulicht dieses Szenario?

A) - Bedrohung 1: Augmentation Data Leak (Leck von Augmentierungsdaten)
- Bedrohung 2: Ausgabe mit klassischer Injection

B) - Bedrohung 1: Manipulation von Augmentierungsdaten
- Bedrohung 2: Augmentation Data Leak

C) - Bedrohung 1: Membership Inference
- Bedrohung 2: Manipulation von Augmentierungsdaten

D) - Bedrohung 1: Ausgabe mit klassischer Injection
- Bedrohung 2: Membership Inference

22 / 40
Allgemeine Governance-Kontrollen sollten implementiert werden, um eine wirksame Sicherheitsaufsicht über KI-Systeme zu gewährleisten.

Welcher Ansatz verdeutlicht dies **am besten**?

A) Die Festlegung klarer Richtlinien, Rollen und Risikomanagementpraktiken als Leitlinien für die sichere Entwicklung, Bereitstellung und Überwachung von KI-Systemen

B) Die Einbindung von Sicherheitskontrollen in die KI-Systeme, um die Security-by-Design-Prinzipien bereits in der Entwurfsphase des KI-Systems anzuwenden

C) Die Einhaltung von Mindestsicherheitsanforderungen, da KI-Systeme standardmäßig sicherer sind und nicht mehrere Sicherheitsebenen erfordern

D) Die Trennung der KI-Governance-Kontrollen von den Informationssicherheitskontrollen, um sicherzustellen, dass diese beiden unterschiedlichen Konzepte gleichermaßen abgedeckt sind

23 / 40
Eine mittelgroße Marketingagentur verfügt über wenig KI-Erfahrung und hat noch keine formelle KI-Governance etabliert. Sie möchte ein generatives KI-Tool zum Verfassen von Kunden-E-Mails einsetzen. Das KI-Tool benötigt Zugriff auf Kundennamen und E-Mail-Adressen. Das Unternehmen sollte vor Nutzung des KI-Tools eine KI-Governance und eine Aufsicht über die KI-Sicherheit einrichten.

Was ist das **absolute Minimum**, das das Unternehmen für die Aufsicht über die KI-Sicherheit unternehmen sollte?

A) Red-Team-Tests an dem E-Mail-System der generativen KI durchführen, um Sicherheitsrisiken vor der Einführung des Systems zu mindern
B) Einen KI-Governance-Spezialisten einstellen, der eine Bedrohungsanalyse durchführt und Richtlinien und Kontrollmaßnahmen für die Einführung des Systems festlegt
C) Eine Bestandsaufnahme der geplanten KI-Nutzung erstellen und eine Risikoanalyse durchführen, um Bedrohungen, Kontrollmaßnahmen und Verantwortlichkeiten zu identifizieren
D) Rechtsberatung von einem Unternehmen einholen, das sich auf KI und die damit verbundenen Anforderungen der KI-Verordnung spezialisiert hat
E) Mit einem Pilotprojekt starten und das KI-E-Mail-System hinsichtlich Leistung, Missbrauch und Vorfällen anhand von Protokollen überwachen

24 / 40
Wie sollte der Geltungsbereich allgemeiner Governance-Kontrollmaßnahmen im Bereich der KI-Sicherheit **am besten** verstanden werden?

A) Als eine Sammlung von Verschlüsselungsmethoden, um die Sicherheit von KI-Ergebnissen nach deren Generierung zu gewährleisten
B) Als eine Reihe übergreifender Kontrollmaßnahmen, die für alle KI-Bedrohungen und Lebenszyklusphasen gelten
C) Als auf die Trainingsphase des Modells beschränkte Kontrollmaßnahmen, die sich auf die Qualität der Datensätze und Optimierungstechniken konzentrieren
D) Als optionale Maßnahmen, die Organisationen ergreifen können, wenn die KI-Systeme keine sensiblen Daten verarbeiten
E) Als technische Schutzmaßnahmen, die ausschließlich während der Runtime angewendet werden, um bösartige Eingaben zu erkennen

25 / 40
Ein Unternehmen im Gesundheitswesen beschließt, für eine Anwendung ein von einem Drittanbieter bereitgestelltes, vorgefertigtes KI-Modell einzusetzen. Die Anwendung verarbeitet Eingaben der Nutzer und ruft externe Daten ab.

Der KI-Sicherheitsingenieur hat Bedenken hinsichtlich der Sicherheit des Modells. Die Aufteilung der Sicherheitsverantwortlichkeiten zwischen dem Anbieter des KI-Modells und dem Unternehmen im Gesundheitswesen wurde nicht in einer Sicherheitsstrategie festgelegt.

Was ist in Bezug auf diese Verantwortlichkeiten **richtig**?

A) Werden vorgefertigte KI-Modelle verwendet, können die meisten Sicherheitskontrollen an den Anbieter delegiert werden, da das Training und das Hosting durch den Anbieter die Sicherheitsrisiken abdecken.
B) Der Betreiber ist für alle Sicherheitskontrollen verantwortlich, wenn er das KI-Modell in seine Systeme integriert und die Sicherheit vor der Bereitstellung berücksichtigt.
C) Der Anbieter hat das KI-Modell trainiert und verwaltet daher die Kontrollmaßnahmen auf Modellebene, während der Betreiber die Kontrollmaßnahmen auf Anwendungsebene verwalten sollte.
D) Der Anbieter ist für alle Sicherheitskontrollen verantwortlich, da er das KI-Modell trainiert hat und es hostet; der Anbieter sollte daher für den Umgang mit allen Risiken und Angriffen verantwortlich sein.

26 / 40
Eine Organisation hat eine kundenorientierte KI-Anwendung bereitgestellt, die ein Large Language Model (LLM) eines Drittanbieters (Third-Party Provider) nutzt. Der Zugriff auf das Modell erfolgt über einen verwalteten API-Dienst (Application Programming Interface). Das Modell wurde von der Organisation nicht weiter trainiert oder modifiziert.

Das Sicherheitsteam stellt fest, dass konstruierte Nutzereingaben dazu führen können, dass das Modell seine konfigurierten Verhaltensgrenzen ignoriert und Ausgaben erzeugt, die gegen die Inhaltsbeschränkungen der Organisation verstoßen. Das Team überlegt, welche Kontrolle implementiert werden sollte, die in die Zuständigkeit der Organisation als Betreiber (Deployer) fällt und nicht in die des Drittanbieters.

Welche Kontrolle sollte das Sicherheitsteam implementieren, um das Problem zu beheben?

A) Eine Ausgabevalidierungsschicht (Output Validation Layer) einfügen, die alle Modellantworten überprüft, bevor diese die Endnutzer erreichen, und alle Ausgaben blockiert, die gegen die Inhaltsrichtlinie der Organisation verstoßen
B) Die Inferenz-Pipeline des Anbieters ändern, um strengere Generierungsbeschränkungen durchzusetzen, die die Erzeugung von Antworten einschränken und das Entstehen unzulässiger Inhalte verhindern
C) Den Anbieter auffordern, das bereitgestellte Modell mithilfe zusätzlicher Daten zur Ablehnung unerlaubter Anfragen und zur Richtlinienkonformität neu zu trainieren und sicherheitsseitig abzustimmen, um seine Widerstandsfähigkeit gegen Jailbreak-Eingaben zu verbessern
D) Vom Anbieter verlangen, dass dieser die Regeln zur Inhaltsmoderation und deren Durchsetzung für den gesamten Dienst überarbeitet und verschärft, damit unsichere Ausgaben vor der Auslieferung global gefiltert werden

27 / 40
Ein Team bereitet Trainingsdaten für einen generativen KI-Assistenten vor. Es stellt fest, dass es nicht alle Felder und Daten benötigt, die derzeit für die Leistung des KI-Modells verwendet werden. Als nächsten Schritt erwägt das Team Kontrollen zur Begrenzung sensibler Daten sowie zur Verbesserung der Vertraulichkeit und Integrität.

Wie lassen sich Vertraulichkeit und Integrität **am besten** verbessern?

A) Alle Felder in den Trainingsdaten beibehalten und die Daten im Ruhezustand sowie in Backups verschlüsseln
B) Identifikatoren in den Trainingsdaten beibehalten, um künftige Aufgaben und Prozesse des Neutrainings zu vereinfachen
C) Felder und Daten entfernen, die für das Modelltraining und die Modellleistung nicht erforderlich sind
D) Ausschließlich synthetische Daten für alle Trainings- und Evaluierungsphasen generativer KI verwenden

28 / 40
Ein Produktteam trainiert ein KI-Modell für den Kundensupport. Der aktuelle Trainingsdatensatz enthält vollständige Kundenadressen, Telefonnummern und Alteinträge, die für das Erreichen der Leistungsziele nicht benötigt werden, künftig aber nützlich sein könnten. Nachdem eine Red-Team-Übung das Risiko von Datenlecks und Manipulation durch memorisierte Details aufgezeigt hat, beschließt das Team, die Offenlegung (Exposure) zu verringern. Dazu entfernt es unnötige Felder und ganze Einträge, sodass der Datensatz keine sensiblen Informationen mehr enthält.

Welche **allgemeine** Kontrolle wurde in dieser Situation zur Begrenzung sensibler Daten implementiert?

A) Zulässige Daten
B) Datenminimierung
C) Trainingsdaten verschleiern
D) Kurze Aufbewahrung

29 / 40
Ein Einzelhandelsunternehmen beschließt, die Zahlungsdaten seiner Kunden nicht mehr zu speichern, sobald eine Transaktion abgeschlossen ist.

Inwiefern **verringert** diese Maßnahme das Risiko einer Datenoffenlegung (Data Exposure)?

A) Sie macht es unnötig, Daten während der Übertragung oder bei ihrer Archivierung zum Zwecke der Auditierbarkeit zu verschlüsseln
B) Sie gewährleistet die Compliance mit allen relevanten Datenschutzvorschriften und verringert die Komplexität
C) Sie gewährleistet, dass unbefugte Nutzer nicht auf das Unternehmensnetzwerk und die Datenbank zugreifen können
D) Sie verhindert, dass im System vorgehaltene Daten durchsickern, rekonstruiert oder abgeleitet werden können

30 / 40
Ein regionales Krankenhausnetzwerk nutzt einen KI-Agenten, um eingehende E-Mails von Lieferanten zu lesen, Antworten zu formulieren und diese automatisch zu versenden. Bei den nach einem kürzlich aufgetretenen Phishing-Sicherheitsvorfall durchgeführten Sicherheitstests stellten die Tester fest, dass eine Prompt-Injection den KI-Agenten dazu bringen könnte, seine Regeln zu ignorieren und unbefugte Antworten zu versenden.

Welche Entwurfsentscheidung eignet sich **am besten**, um die Auswirkungen dieses unerwünschten Verhaltens zu begrenzen?

A) Alle Integrationen von KI-Agenten deaktivieren und die meisten Prozesse wieder auf vollständig manuelle Arbeitsabläufe umstellen
B) Die Schulungen der Nutzer zur Prompt-Sicherheit erweitern und die Notwendigkeit der Aufsicht beim agentengestützten E-Mail-Versand betonen
C) Weitreichende autonome Versandberechtigungen beibehalten und auf ein fortschrittlicheres generatives Modell upgraden
D) Die Berechtigungen des KI-Agenten auf aufgabenspezifische Bereiche beschränken und für Hochrisiko-Aktionen eine Genehmigung fordern

31 / 40
Wie lassen sich die Auswirkungen unerwünschten Modellverhaltens **am besten** auf das absolute Minimum beschränken?

A) Indem man Prompt-Hardening, rollenbasierten Zugriff und geplante Audits der Modellausgaben hervorhebt
B) Indem man sich auf automatisierte Ausgabefilter sowie die Protokollierung von Ausnahmen verlässt und gelegentliche Stichprobenprüfungen bezüglich Fairness durchführt
C) Indem man auf unerwünschten Bias testet und Kontrollen wie Aufsicht, Least Privilege und kontinuierliche Validierung anwendet
D) Indem man eine strenge Eingabevalidierung, eingeschränkte Tool-Berechtigungen und regelmäßige Human-in-the-Loop-Überprüfungen nutzt

32 / 40
Die Begrenzung unerwünschten Modellverhaltens ist eine entscheidende Strategie zur Begrenzung der Auswirkungen und zur Kontrolle des Auswirkungsradius.

Wie **verbessert** diese Strategie die Leistung und **verringert** das Risiko?

A) Sie konzentriert sich auf Angriffe, die KI-Systeme beeinträchtigen können, und beschränkt die Aufsicht von Systemleistungsfaktoren und Genauigkeit
B) Sie implementiert Kontrollen, die kontinuierliche Tests und Validierungen mit der Zeit überflüssig machen und so die Komplexität reduzieren
C) Sie verbessert die betriebliche Zuverlässigkeit und die Ressourceneffizienz und minimiert gleichzeitig Sicherheitsvorfälle und unnötigen Rechenaufwand durch Anfragen
D) Sie erhöht die Modellautonomie, sodass das Modell mehr Ausgaben generieren kann, um die Systemleistung kontinuierlich zu verbessern

33 / 40
Was ist im größeren Kontext des KI-Red-Teamings die **Hauptaufgabe** von KI-Sicherheitstests?

A) Die Beurteilung, ob ein KI-Modell widerstandsfähig gegenüber bestimmten Angriffen ist, indem diese Angriffe in einer kontrollierten Umgebung reproduziert werden
B) Die Bewertung, ob ein KI-Modell genaue und rückverfolgbare Ausgaben liefert, wenn es mit den erwarteten Eingabedaten konfrontiert wird
C) Die Messung, ob ein KI-Modell bei der Verarbeitung eines großen Anfrageaufkommens die erwartete Recheneffizienz aufweist
D) Die Verifizierung mittels Standardverfahren der Qualitätssicherung, dass ein KI-Modell die festgelegten geschäftlichen Anforderungen erfüllt

34 / 40
Ein Unternehmen stellt einen generativen KI-Assistenten bereit und möchte KI-Sicherheitstests definieren, die über herkömmliche Penetrationstests hinausgehen. Im ersten Schritt identifiziert das Testteam die Bedrohungen, auf die getestet werden sollte.

Welche **spezifisch** mit generativen KI-Systemen verbundenen Bedrohungen sollten in die KI-Sicherheitstests einbezogen werden?

A) Evasion Attacks (Umgehungsangriffe), unsichere Ausgabeverarbeitung und Model Poisoning (Modellvergiftung)
B) Model Exfiltration (Modellausleitung), Prompt-Injection und Evasion Attacks
C) Prompt-Injection, Offenlegung sensibler Daten und unsichere Ausgabeverarbeitung
D) Ausgabe sensibler Daten durch das Modell, Modelldrift und unsicherer Tool-Aufruf

35 / 40
Ein KI-Sicherheitsteam testet einen generativen KI-Chatbot auf seine Widerstandsfähigkeit gegen Prompt-Injection. Während des Tests hat das System mehrere konstruierte Angriffseingaben blockiert.

Welcher Schritt wird als **Nächstes** empfohlen, um die Sicherheitsaufsicht des generativen KI-Chatbots aufrechtzuerhalten?

A) Die Eingaben beispielsweise durch den Austausch von Synonymen, Änderung der Kodierung oder Anpassung der Formatierung variieren, um zu versuchen, die Erkennung durch das KI-System zu umgehen
B) Den aktuellen Testzyklus abschließen und die Abwehrmaßnahmen als ausreichend erachten, da das System die meisten Prompt-Injection-Versuche blockiert hat
C) Denselben Test mehrmals wiederholen, um die Stichprobengröße zu erhöhen und statistisch aussagekräftigere Daten bezüglich der Abwehrleistung zu erhalten
D) Die Eingaben für Prompt-Injection-Angriffe direkt an das Modell senden, um es hinsichtlich ähnlicher Bedrohungen in zukünftigen Bereitstellungen besser zu trainieren

36 / 40
Das KI-System eines Unternehmens stützt sich beim Training und bei der Entscheidungsfindung auf große Mengen personenbezogener Daten.

Um welches Risiko sollte sich das Unternehmen aus ethischer Sicht **vorrangig** kümmern?

A) Die Erhebung, Speicherung und Weiterverwendung von Daten, die über den gerechtfertigten Zweck hinausgeht
B) Die Komplexität der Benutzeroberfläche, die die Überprüfung automatisierter Entscheidungen erschwert
C) Die sinkende Genauigkeit des Modells bei ungleichmäßiger Ressourcenzuweisung
D) Softwareupdates, die plattformübergreifende Kompatibilitätsprobleme verursachen

37 / 40
Eine Online-Buchhandlung analysiert mit Hilfe eines KI-Systems Browserverlauf, Kauftrends und Vorbestellungen, um die Nachfrage vorherzusagen und sicherzustellen, dass beliebte Bücher rechtzeitig vorrätig sind. Später verwendet das Unternehmen dieselben Kundendaten erneut, um ein separates KI-Modell zu trainieren, das detaillierte Marketingprofile erstellt und gezielte Werbung an einzelne Käufer sendet.

Gegen welchen Datenschutzgrundsatz verstößt dies **am unmittelbarsten**?

A) Datenminimierung
B) Begrenzung der Datenaufbewahrung
C) Transparenz und Einwilligung
D) Verwendungsbegrenzung und Zweckbestimmung

38 / 40
Eine Organisation möchte ein Framework erstellen, um den verantwortungsvollen KI-Einsatz in allen ihren Geschäftsbereichen zu regeln.

Sie nutzt die Norm ISO/IEC 27001 für das Informationssicherheitsmanagement. Das Sicherheitsteam möchte seine KI-Governance anhand einer ISO/IEC-Norm aufbauen, die für die KI-Governance denselben strukturellen Zweck erfüllt wie die Norm ISO/IEC 27001 für die Informationssicherheit.

Um welche Norm handelt es sich hierbei?

A) ISO/IEC 23894
B) ISO/IEC 27005
C) ISO/IEC 42001
D) ISO/IEC 5338

39 / 40
Ein Unternehmen möchte ein KI-System einsetzen, um Personen im Bewerbungsverfahren anhand ihrer Lebensläufe automatisch zu prüfen und in eine Rangfolge zu bringen.

In welche Kategorie ist die Nutzung dieses KI-Systems laut KI-Verordnung einzuordnen?

A) Unannehmbares Risiko
B) Hohes Risiko
C) Begrenztes Risiko
D) Minimales oder kein Risiko

40 / 40
Ein KI-Entwicklungsteam bereitet die Trainingsdaten für ein neues Modell für maschinelles Lernen vor. Der Teamleiter möchte sicherstellen, dass das Vorgehen bei der Datenbeschaffung die Compliance mit dem Urheberrecht (Copyright Compliance) unterstützt.

Welche Vorgehensweise bei der Datenbeschaffung ist **am ethischsten**?

A) Daten ausschließlich aus öffentlich zugänglichen Quellen zu erheben
B) Trainingsdaten ausschließlich organisationsintern zu erstellen
C) Daten mit stillschweigender Genehmigung aus Quellen herunterzuladen
D) Ausschließlich Daten zu verwenden, die wiederholt online weiterverbreitet wurden

Antwortschlüssel

1 / 40
Ein IT-Sicherheitsmanager ist bei einem Logistikunternehmen tätig. Das Unternehmen hat kürzlich beschlossen, KI-Tools zur Optimierung seiner Lieferrouten einzusetzen. Der IT-Sicherheitsmanager ist für die Organisation der KI-Sicherheit verantwortlich. Die ersten beiden Schritte des G.U.A.R.D.-Frameworks sind bereits abgeschlossen:

- Der erste Schritt Govern (Steuern) wurde durch die Erstellung einer KI-Sicherheitsrichtlinie umgesetzt.
- In der zweiten Stufe Understand (Verstehen) wurden alle KI-Assets und die damit verbundenen Risiken erfasst.

Was ist der **nächste** Schritt, den der IT-Sicherheitsmanager unternehmen sollte?

A) Anpassung der Sicherheitspraktiken um KI-spezifische Bedrohungsmodellierung, Tests und Kontrollen der Lieferkette
B) Analyse der identifizierten KI-Risiken und Festlegung einer Strategie zur Risikopriorisierung entsprechend dem potenziellen Schadensausmaß
C) Anwendung von Sicherheitskontrollen auf alle KI-Systeme zur Gewährleistung einer verantwortungsvollen KI-Sicherheit entlang des gesamten KI-Lebenszyklus
D) Auswertung der im Schritt Govern (Steuern) erstellten KI-Sicherheitsrichtlinie, um die erfassten KI-Assets und deren Risiken einzubeziehen

A) Richtig. Nach den Schritten Govern (Steuern) und Understand (Verstehen) folgt im G.U.A.R.D.-Framework als nächster Schritt Adapt (Anpassen). Er hat zum Ziel, die Sicherheitspraktiken zu aktualisieren, um KI-spezifische Bedrohungen, Sicherheitstests und Maßnahmen zur Lieferkette einzubeziehen. (Literatur: A, Kapitel 0)
B) Falsch. Dies gehört zu dem bereits abgeschlossenen Schritt Understand (Verstehen).
C) Falsch. Die pauschale Anwendung von Kontrollmaßnahmen ohne Berücksichtigung des Kontexts entspricht nicht dem G.U.A.R.D.-Framework. Im nächsten Schritt werden die Praktiken angepasst und nicht wahllos Kontrollmaßnahmen angewendet. Die Anwendung von Kontrollmaßnahmen erfolgt in einer späteren Phase.
D) Falsch. Die Überprüfung der Richtlinien erfolgt im Schritt Govern (Steuern). Auf den Schritt Understand (Verstehen) folgt der Schritt Adapt (Anpassen), nicht die Überarbeitung der Governance.

2 / 40
Ein Unternehmen überprüft seine KI-Strategie, um sicherzustellen, dass diese einen geschäftlichen Mehrwert schafft und gleichzeitig ethischen sowie operativen Anforderungen gerecht wird. Der KI-Sicherheitsbeauftragte unterscheidet dabei zwischen den beiden zentralen Dimensionen verantwortungsvolle KI und vertrauenswürdige KI.

Was ist der **wesentliche** Unterschied zwischen verantwortungsvoller KI und vertrauenswürdiger KI?

A) - Verantwortungsvolle KI konzentriert sich auf Ethik, Gesellschaft und Governance.
- Vertrauenswürdige KI konzentriert sich auf technische und betriebliche Aspekte, Transparenz und Erklärbarkeit.

B) - Verantwortungsvolle KI konzentriert sich auf die Compliance mit einschlägigen Gesetzen und Rechtsvorschriften.
- Vertrauenswürdige KI konzentriert sich auf Datenschutz und Cybersicherheit, um den Datenschutzgrundsätzen zu entsprechen.

C) - Verantwortungsvolle KI konzentriert sich auf die Systemgenauigkeit und technische Leistungskennzahlen.
- Vertrauenswürdige KI konzentriert sich auf ethische Leitlinien und gesellschaftliche Auswirkungen, um das Vertrauen der Verbraucher zu gewinnen.

D) - Verantwortungsvolle KI konzentriert sich auf die Anbieterauswahl, Risiken und das Kostenmanagement.
- Vertrauenswürdige KI konzentriert sich auf eine sichere und zuverlässige Schulung der Nutzer sowie die Einführung von KI-Systemen.

A) Richtig. Verantwortungsvolle KI legt den Schwerpunkt auf Ethik, Gesellschaft und Governance (Grundsätze, Richtlinien, Aufsicht), während vertrauenswürdige KI die eher technischen und betrieblichen Aspekte des KI-Einsatzes in einer Organisation in den Vordergrund stellt, wie beispielsweise Robustheit, Zuverlässigkeit, Transparenz und Erklärbarkeit. (Literatur: A, Kapitel 0)

B) Falsch. Dies reduziert verantwortungsvolle KI auf die regulatorische Compliance und vertrauenswürdige KI auf Privatsphäre und Cybersicherheit. Dabei handelt es sich lediglich um Teilbereiche, die den umfassenderen Unterschied zwischen den beiden Konzepten nicht vollständig abbilden.

C) Falsch. Dies kehrt den Schwerpunkt um: Genauigkeit und Leistung sind Teil der technischen Vertrauenswürdigkeit, während Ethik und gesellschaftliche Auswirkungen für verantwortungsvolle KI von zentraler Bedeutung sind.

D) Falsch. Die Auswahl von Anbietern, Kosten und Schulungen für Nutzer sind Themen des Prozesses Change Management, nicht der zentrale konzeptionelle Unterschied zwischen verantwortungsvoller und vertrauenswürdiger KI.

3 / 40
Ein Cybersicherheitsteam verfügt bereits über ein Framework zur Erkennung von Bedrohungen. Das Team muss nun auch ein neues Large Language Model (LLM) schützen, das im Kundensupport für einen Chatbot eingesetzt wird.

Das Team möchte dieselben Tools verwenden, die es auch für Webanwendungen einsetzt:
- Signaturbasierte Erkennungsregeln
- Feste Kontrollen zur Eingabevalidierung

Warum reicht dieser Ansatz **nicht** aus, um das LLM zu sichern?

A) Weil KI-Systeme andere Angriffsflächen aufweisen und spezifische Kontrollen für das Verhalten des LLM sowie die Datenpipeline erfordern
B) Weil das Framework zur Erkennung von Bedrohungen keine Verfahren zur Reaktion auf Sicherheitsvorfälle umfasst und daher für LLMs ungeeignet ist
C) Weil LLMs Ausgaben generieren, die zu variabel sind, um von regelbasierten Systemen bewertet werden zu können, und daher von einem Menschen überprüft werden müssen
D) Weil LLMs Anfragen schneller verarbeiten als Webanwendungen, sodass signaturbasierte Regeln nicht mit ihnen Schritt halten können
E) Weil Firewalls für Webanwendungen (Web Application Firewall, WAF) nicht für den Einsatz mit LLMs lizenziert sind und separate Vereinbarungen mit den Anbietern erfordern

A) Richtig. Herkömmliche Cybersicherheitstools wie WAFs und signaturbasierte Erkennungssysteme sind darauf ausgelegt, bekannte Angriffsmuster gegen feste Software-Schnittstellen zu identifizieren. KI-Systeme führen völlig neue Angriffsflächen ein: Die Trainingsdaten können manipuliert werden, das Modell selbst kann umgedreht werden, um sensible Informationen zu extrahieren, und die Ausgaben können durch adversariale Eingaben manipuliert werden, die keiner bekannten Signatur entsprechen. Diese Bedrohungen erfordern KI-spezifische Kontrollmaßnahmen, die nicht nur auf die Netzwerkschnittstelle, sondern auch auf das Verhalten des Modells und die Datenpipeline einwirken. (Literatur: A, Kapitel 0 und 1)
B) Falsch. Herkömmliche Frameworks für Cybersicherheit umfassen durchaus Verfahren zur Reaktion auf Vorfälle. Diese Aussage stellt die Einschränkung herkömmlicher Sicherheitsansätze für KI-Systeme jedoch falsch dar: Das Problem liegt in der Abdeckung der Angriffsfläche und nicht in fehlenden Verfahren.
C) Falsch. Auch wenn die Ausgaben von LLMs sehr variabel sind, können herkömmliche regelbasierte Kontrollmaßnahmen dennoch zu einer mehrschichtigen Verteidigung in KI-Umgebungen beitragen. Die Einschränkung besteht darin, dass sie KI-spezifische Angriffsflächen wie Modellmanipulation, adversariale Eingaben, Prompt-Injection und Angriffe auf Trainingsdaten nicht angemessen berücksichtigen.
D) Falsch. Die Verarbeitungsgeschwindigkeit ist ein leistungsbezogener Aspekt und nicht der Grund dafür, dass die signaturbasierte Erkennung für KI-Systeme unzureichend ist. Die Einschränkung liegt hier nicht im Zeitfaktor, sondern in der Art der Angriffsfläche.
E) Falsch. Es gibt keine allgemeinen Lizenzbeschränkungen, die den Einsatz von WAFs an Endpunkten von KI-Modellen verhindern würden. Diese Option beschreibt eher einen beschaffungstechnischen Aspekt als eine wesentliche sicherheitstechnische Einschränkung des vorgeschlagenen Ansatzes.

4 / 40
Bei der Nutzung von KI verfügt ein Unternehmen über KI-Assets. Für jedes KI-Asset lassen sich spezifische Hauptbedrohungen identifizieren.

Welche Kombination aus einem KI-Asset und seiner Hauptbedrohung ist **korrekt**?

A) - Asset: Von Netzwerk-Firewalls generierte Audit-Protokolle
- Bedrohung: AI Resource Exhaustion (KI-Ressourcenerschöpfung)

B) - Asset: Authentifizierungsdaten von Nutzern
- Bedrohung: Die Ausgabe enthält eine klassische Injection

C) - Asset: In einem Versionskontrollsystem gespeicherter Software-Quellcode
- Bedrohung: Direct Runtime Model Poisoning (direkte Modellvergiftung während der Runtime)

D) - Asset: Trainingsdaten, die zum Aufbau eines Modells verwendet werden
- Bedrohung: Offenlegung sensibler Daten in der Ausgabe

A) Falsch. Audit-Protokolle von Netzwerk-Firewalls sind ein allgemeines IT-Sicherheits-Asset und kein KI-Asset, wie es in der OWASP AI Exchange beschrieben wird. AI Resource Exhaustion bezieht sich in der Regel auf die übermäßige Nutzung von Rechenressourcen, die für KI-Aufgaben benötigt werden, was keinen direkten Bezug zu Audit-Protokollen hat.

B) Falsch. Authentifizierungsdaten von Nutzern stellen ein gängiges IT-Sicherheitsrisiko dar, nicht jedoch ein KI-Asset, wie es in der OWASP AI Exchange beschrieben wird. Herkömmliche Injektionsbedrohungen wie SQL- oder Befehlsinjektionen sind eher für die Anwendungssicherheit relevant als für KI-Ausgaben. Diese Zuordnung ist nicht korrekt.

C) Falsch. Software-Quellcode kann zwar ein KI-Asset sein, wenn er sich auf KI-Modelle bezieht, aber das Direct Runtime Model Poisoning bezieht sich auf die Manipulation eines KI-Modells während dessen Ausführung und nicht auf dessen Quellcode in der Versionskontrolle.

D) Richtig. Trainingsdaten sind ein kritisches KI-Asset. Die Offenlegung von Daten in den Ausgaben stellt eine ernstzunehmende Bedrohung dar, insbesondere wenn sensible oder geschützte Daten aus dem Trainingsdatensatz versehentlich in den von der generativen KI generierten Ausgabedaten erscheinen. Diese Kombination identifiziert das Asset und die damit verbundene Bedrohung im KI-Kontext korrekt. (Literatur: A, Kapitel 0)

5 / 40
Eine Anwaltskanzlei nutzt zur Dokumentenprüfung und -analyse ein KI-System. Das Sicherheitsteam arbeitet mit einem Risikomanagementansatz. Eine Bestandsaufnahme aller potenziellen Bedrohungen für das KI-System wurde durchgeführt.

Welchen Schritt sollte das Team als **Nächstes** unternehmen?

A) Einen Prozess der Bedrohungsmodellierung anwenden, um die Bedrohungen in eine Reihe konkreter und priorisierter Risiken zu überführen.
B) Jeder identifizierten Bedrohung einen Risikoverantwortlichen zuweisen, um von Anfang an „Protection by Design“ zu gewährleisten.
C) Für alle Bedrohungen Sicherheitskontrollen implementieren, um weitere Risiken entlang des gesamten Lebenszyklus des KI-Systems zu mindern.
D) Die vollständige Liste der Bedrohungen an die relevanten Stakeholder weitergeben, um das Bewusstsein zu schärfen und Transparenz zu schaffen.

A) Richtig. Nach Auflistung der Bedrohungen sollte das Team einen Prozess der Bedrohungsmodellierung anwenden, der die aufgelisteten Bedrohungen mit einer Reihe konkreter, priorisierter Risiken in Verbindung setzt. Dies ermöglicht der Organisation, sich auf die Anwendung von Kontrollmaßnahmen für die bedeutendsten Risiken zu konzentrieren. (Literatur: A, Kapitel 0)
B) Falsch. Die Zuweisung von Risikoverantwortlichen vor der Analyse und Priorisierung der Bedrohungen überspringt einen entscheidenden Schritt. Ohne ein Verständnis der konkreten Auswirkungen und der Eintrittswahrscheinlichkeit der einzelnen Bedrohungen fehlt den Verantwortlichkeitszuweisungen ein aussagekräftiger Kontext.
C) Falsch. Sicherheitskontrollen für alle Bedrohungen ohne vorherige Priorisierung sind ineffizient und können Ressourcen fehlleiten. Bevor Kontrollmaßnahmen ausgewählt und umgesetzt werden, sollten Bedrohungen daher zunächst durch eine Bedrohungsmodellierung priorisiert werden.
D) Falsch. Die Weitergabe einer unbearbeiteten, nicht priorisierten Bedrohungsliste hilft den Stakeholdern nicht dabei, fundierte Entscheidungen zu treffen. Der Prozess der Bedrohungsmodellierung muss die Liste zunächst in konkrete, priorisierte Risiken überführen, bevor eine Kommunikation sinnvoll ist.

6 / 40
Ein KI-Sicherheitsingenieur überprüft das Design der Zugriffskontrolle für ein KI-Multi-Agenten-System. Alle Agenten im System teilen sich ein Dienstkonto. Dieses gemeinsame Dienstkonto hat Lese- und Schreibzugriff auf die Produktionsdatenbank, das interne Dateisystem und das externe API-Gateway (Application Programming Interface).

Die verschiedenen KI-Agenten rufen autonom Tools auf und führen mehrstufige Workflows über interne Systeme hinweg aus. Die Bereitstellung erfolgte über ein gemeinsames Dienstkonto, um die Verwaltung der Anmeldedaten entlang der gesamten KI-Agent-Pipeline zu vereinfachen.

Warum ist dieses Design mit einem gemeinsamen Dienstkonto bei einer agentischen KI **besonders riskant**?

A) Weil KI-Agenten ein höheres Volumen an API-Aufrufen generieren als menschliche Nutzer, sodass es wahrscheinlicher ist, dass Ratenbegrenzungsmaßnahmen ausgelöst werden und das gemeinsam genutzte Dienstkonto gesperrt wird
B) Weil KI-Agenten keine Multi-Faktor-Authentifizierung verwenden, sodass die gemeinsam genutzten Anmeldedaten einen Single Point of Failure ohne weitere Verifizierungsebenen darstellen, den ein Angreifer ausnutzen kann
C) Weil KI-Agenten innerhalb einer einzigen Sitzung autonom über mehrere Schlussfolgerungsschritte hinweg agieren, sodass ein kompromittierter Agent Aktionen über alle zugänglichen Systeme hinweg verketten kann
D) Weil regulatorische Rahmenbedingungen für KI-Systeme vorschreiben, dass jeder KI-Agent mit eindeutigen Anmeldedaten ausgestattet sein muss, sodass gemeinsam genutzte Dienstkonten einen direkten Compliance-Verstoß darstellen
E) Weil die gemeinsame Nutzung von Dienstkonten durch mehrere KI-Agenten während der vorgeschriebenen Rotation der Anmeldedaten und bei Aktivitäten im Zugriffsmanagement zu einer zu hohen betrieblichen Komplexität führt

A) Falsch. Das Volumen der API-Aufrufe ist ein Problem hinsichtlich Leistung und Verfügbarkeit, kein Sicherheitsrisiko, das spezifisch mit dem Konzept gemeinsamer Anmeldedaten verbunden ist. Die Ratenbegrenzung betrifft alle Kontotypen und erklärt nicht, warum ein gemeinsames Dienstkonto für Agenten im Vergleich zu menschlichen Nutzern, die dasselbe Dienstkonto nutzen, ein unverhältnismäßig höheres Risiko darstellt.
B) Falsch. Das Fehlen einer Multi-Faktor-Authentifizierung für Dienstkonten ist ein allgemeines Sicherheitsproblem, das sowohl für gemeinsame Dienstkonten menschlicher Nutzer als auch für gemeinsame Dienstkonten von Agenten gilt. Es erklärt nicht, warum das Risiko speziell im Kontext einer agentischen KI unverhältnismäßig höher ist.
C) Richtig. Das in der Frage angesprochene Sicherheitsrisiko steht im Zusammenhang mit übermäßiger Handlungsautonomie. Agentische KI-Systeme können innerhalb einer einzigen Sitzung autonom Aktionen über mehrere Systeme, Tools und Schlussfolgerungsschritte hinweg verketten. Ein kompromittierter Agent, der ein gemeinsames Dienstkonto nutzt, kann daher umfangreiche unbefugte Aktionen ausführen, bevor ein menschliches Eingreifen oder eine Überprüfung erfolgt. (Literatur: A, Kapitel 0 und 1)
D) Falsch. Die regulatorischen Anforderungen an die Bereitstellung von Anmeldedaten für KI-Agenten variieren je nach Rechtsraum und Branche und sind auf der beschriebenen Detailebene nicht einheitlich festgelegt. Diese Antwort geht nicht auf das in der Frage dargestellte Sicherheitsrisiko ein.
E) Falsch. Hier wird eine betriebliche Herausforderung angesprochen, nicht das spezifische Sicherheitsrisiko, das durch die Gestaltung gemeinsamer Dienstkonten im Kontext von agentischer KI entsteht.

7 / 40
Ein Ingenieur für maschinelles Lernen bei einem Finanzinstitut untersucht zwei separate Vorfälle im Zusammenhang mit dem Betrugserkennungsmodell des Instituts.

- **Vorfall 1**: Ein externer Angreifer, der keinen Zugriff auf die Modellarchitektur, die Trainingsdaten oder die Parameter des Modells hat, übermittelt Tausende von manipulierten Transaktionen. Diese werden iterativ verfeinert, je nachdem, ob der Betrugsalarm ausgelöst wurde oder nicht.
- **Vorfall 2**: Ein Forscher verwendet ein lokal trainiertes Ersatzmodell mit ähnlicher Modellarchitektur, um adversariale Transaktionen zu generieren. Diese umgehen das Produktionsmodell erfolgreich, ohne dass zuvor eine Interaktion stattgefunden hat.

Um welche zwei Arten von Evasion Attacks (Umgehungsangriffe) handelt es sich hier?

A) - Vorfall 1: Evasion After Poisoning (Umgehung nach Vergiftung)
- Vorfall 2: Perfect-Knowledge Evasion (Umgehungsangriff mit vollständigem Wissen)

B) - Vorfall 1: Partial-Knowledge Evasion (Umgehungsangriff mit Teilwissen)
- Vorfall 2: Transfer Attack (Transferangriff)

C) - Vorfall 1: Perfect-Knowledge Evasion
- Vorfall 2: Zero-Knowledge Evasion (Umgehungsangriff mit Nullwissen)

D) - Vorfall 1: Transfer Attack
- Vorfall 2: Partial-Knowledge Evasion

E) - Vorfall 1: Zero-Knowledge Evasion
- Vorfall 2: Transfer Attack

A) Falsch. Vorfall 1 wird fälschlicherweise als Evasion After Poisoning identifiziert. Es gibt keine Hinweise darauf, dass der Angreifer die Trainingsdaten verändert hat. Der Angreifer hat lediglich zahlreiche Transaktionen getestet und aus den Antworten des Modells gelernt. Vorfall 2 wird fälschlicherweise als Perfect-Knowledge Evasion identifiziert. Der Forscher hatte keinen vollständigen Zugriff auf das eigentliche Modell. Er verwendete ein Ersatzmodell.

B) Falsch. Vorfall 1 wird fälschlicherweise als Partial-Knowledge Evasion identifiziert. Der Angreifer hatte keinen Zugriff auf den Aufbau, die Daten oder die Parameter des Modells, sondern nutzte lediglich dessen Ausgaben. Vorfall 2 wird korrekt als Transfer Attack identifiziert. Die Gesamtantwort ist dennoch falsch, da es sich bei Vorfall 1 nicht um eine Partial-Knowledge Evasion handelt.

C) Falsch. Vorfall 1 wird fälschlicherweise als Perfect-Knowledge Evasion identifiziert. Dieser Angriff gewährt dem Angreifer keinen vollständigen Zugriff auf das Modell. Vorfall 2 wird fälschlicherweise als Zero-Knowledge Evasion identifiziert. Der Forscher verwendete ein Ersatzmodell, um den Angriff zu entwickeln; es handelt sich also um eine Transfer Attack.

D) Falsch. Vorfall 1 wird fälschlicherweise als Transfer Attack identifiziert. In Vorfall 1 hat der Angreifer das echte Modell mehrfach getestet und die Eingaben anhand der Antworten angepasst. Vorfall 2 wird fälschlicherweise als Partial-Knowledge Evasion identifiziert. Der Forscher hat ein separates Modell verwendet, daher handelt es sich um eine Transfer Attack.

E) Richtig. Vorfall 1 wird korrekt als Zero-Knowledge Evasion identifiziert. Der Angreifer hatte keinen internen Zugriff und nutzte ausschließlich die Ausgaben des Modells. Vorfall 2 wird korrekt als Transfer Attack identifiziert. Der Forscher baute den Angriff auf einem Ersatzmodell auf, und dieser täuschte auch das echte Modell. (Literatur: A, Kapitel 2)

8 / 40
Erick ist Nutzer einer Social-Media-Plattform, die Kommentare mithilfe eines KI-Systems automatisch moderiert. Er versucht, einen gegen die Richtlinien verstoßenden Kommentar an der automatisierten Moderation der Plattform vorbeizuschleusen, da er die Interaktion mit seinem Beitrag steigern möchte. Erick hat keinerlei Informationen über die Architektur, die Parameter oder die Trainingsdaten des Modells und sieht beim Versuch, seinen Kommentar zu veröffentlichen, lediglich eine binäre Meldung: entweder „blockiert“ oder „veröffentlicht“. Indem er wiederholt kleine Änderungen am Wortlaut seines Kommentars vornimmt, findet Erick schließlich eine Version, die veröffentlicht wird.

Um welche Art von Evasion Attack (Umgehungsangriff) handelt es sich hier?

A) Zero-Knowledge Evasion (Umgehungsangriff mit Nullwissen)
B) Partial-Knowledge Evasion (Umgehungsangriff mit Teilwissen)
C) Perfect-Knowledge Evasion (Umgehungsangriff mit vollständigem Wissen)
D) Evasion After Poisoning (Umgehung nach Vergiftung)
E) Transfer Attack (Transferangriff)

A) Richtig. Erick verfügt über keinerlei Informationen zum KI-Modell und nutzt eine abfragebasierte Angriffsstrategie, um eine erfolgreiche Angriffseingabe zu finden, wodurch er seinen Kommentar veröffentlichen kann. Dies entspricht der Definition einer Zero-Knowledge Evasion. (Literatur: A, Kapitel 2)
B) Falsch. Eine Partial-Knowledge Evasion würde einige Zusatzinformationen erfordern, um den Angriff zu lenken. Über diese verfügt Erick jedoch nicht.
C) Falsch. Eine Perfect-Knowledge Evasion würde den vollständigen Zugriff auf die Modellarchitektur, die Parameter und die Trainingsdaten des KI-Modells erfordern, um Eingaben präzise zu gestalten.
D) Falsch. Eine Evasion After Poisoning bezieht sich auf eine Evasion Attack, die durch die vorherige Vergiftung der Trainingsdaten des KI-Modells ermöglicht wird. Dies ist hier nicht der Fall.
E) Falsch. Bei einer Transfer Attack werden adversariale Beispiele mithilfe eines Ersatzmodells erstellt und anschließend auf das Zielmodell angewendet.

9 / 40
Ein Rechtsdienstleistungsunternehmen entwickelt einen Chatbot, der Fragen von Nutzern beantwortet und Informationen aus internen und externen Quellen wie Websites und Unternehmensdokumenten abrufen kann. Während einer Demo löst das Verhalten des Chatbots eine Untersuchung durch die KI-Sicherheit aus. Sie identifiziert zwei separate Bedrohungen:

- **Bedrohung 1:** Eine Nachricht, die so gestaltet ist, dass sie die Anweisungen und Sicherheitsregeln des Chatbots außer Kraft setzt und dazu führen kann, dass der Chatbot den Befehlen des Testers folgt, anstatt sich an die konfigurierten Richtlinien zu halten.
- **Bedrohung 2:** Eine vom Chatbot abgerufene Anbieter-Website enthält eingebetteten Text mit Anweisungscharakter. Wenn der Chatbot diese Seite aufruft, behandelt er den Inhalt als Befehle und weicht von seinen Richtlinien ab.

Um welche Arten von Bedrohungen handelt es sich?

A) - Bedrohung 1: Evasion Attack (Umgehungsangriff)
- Bedrohung 2: Data Poisoning (Datenvergiftung)

B) - Bedrohung 1: Data Poisoning
- Bedrohung 2: direkte Prompt-Injection

C) - Bedrohung 1: direkte Prompt-Injection
- Bedrohung 2: indirekte Prompt-Injection

D) - Bedrohung 1: indirekte Prompt-Injection
- Bedrohung 2: Evasion Attack

A) Falsch. Bedrohung 1 ist keine Evasion Attack, da das Modell nicht von einer adversarialen Eingabe in die Irre geführt wurde. Bedrohung 2 ist kein Data Poisoning, da die Trainingsdaten des Modells nicht verändert wurden.

B) Falsch. Bedrohung 1 ist kein Data Poisoning, da die Trainingsdaten des Modells nicht verändert wurden. Bedrohung 2 ist keine direkte Prompt-Injection, da die Anweisungen indirekt aus Webinhalten stammten.

C) Richtig. Bedrohung 1 ist eine direkte Prompt-Injection, bei der Nutzer-Prompts die KI täuschen. Bedrohung 2 ist eine indirekte Prompt-Injection, bei der Webinhalte versteckte Anweisungen enthalten. (Literatur: A, Kapitel 2)

D) Falsch. Bedrohung 1 ist keine indirekte Prompt-Injection, da die Anweisungen direkt vom Nutzer stammten. Bedrohung 2 ist keine Evasion Attack, da das Modell nicht durch Eingabestörungen getäuscht wurde.

10 / 40
Ein Beratungsunternehmen setzt zur Unterstützung bei alltäglichen Aufgaben ein KI-System ein. Nach einem Gespräch mit einem Kunden fasst der KI-Assistent das Gespräch zusammen. Der KI-Assistent beginnt, vertrauliche interne Links in den Entwurf einer E-Mail einzufügen, die an den Kunden gesendet werden soll. Ein autonomer Anomaliedetektor unterbricht jedoch den Versand und stellt den Entwurf unter Quarantäne.

Der KI-Sicherheitsarchitekt möchte den Zugriff des KI-Besprechungsassistenten vorab auf die Berechtigungen der betreuten Personen beschränken.

Welche Schichten des Prompt-Injection-Schutzes werden hier dargestellt?

A) Automatisierte Aufsicht und benutzerbasiertes Prinzip der geringsten Berechtigungen
B) Just-in-Time-Autorisierung und menschliche Aufsicht
C) Modellabgleich und absichtsbasiertes Prinzip der geringsten Berechtigungen
D) Verarbeitung von Eingaben und Ausgaben bei Prompt-Injection sowie Ratenbegrenzung

A) Richtig. Bei einem autonomen Anomaliedetektor, der den Entwurf unterbricht und unter Quarantäne stellt, handelt es sich um automatisierte Aufsicht. Dem Assistenten nur die vorab zugewiesenen Berechtigungen der betreuten Personen zu gewähren entspricht dem benutzerbasierten Prinzip der geringsten Berechtigungen. (Literatur: A, Kapitel 2)
B) Falsch. Eine Eskalation oder Genehmigung pro Aktion, die eine Just-in-Time-Autorisierung rechtfertigen würde, gibt es nicht. Die Aufsicht erfolgt automatisiert, nicht durch Menschen.
C) Falsch. Das Szenario erwähnt keine Anpassung des Modells und das beschriebene Prinzip der geringsten Berechtigungen basiert auf der Identität des Nutzers und vorab zugewiesenen Rechten, nicht auf dynamischen, absichtsbezogenen Berechtigungen.
D) Falsch. Eine Ratenbegrenzung wird nicht beschrieben und bei der dargestellten Kontrolle handelt es sich um einen selbstständig arbeitenden Anomaliedetektor (automatisierte Aufsicht), nicht um spezifische Maßnahmen zur Kennzeichnung von Ein- und Ausgabedaten.

11 / 40
Ein Prüfer bewertet die Datenschutzrisiken eines im Gesundheitswesen eingesetzten Large Language Models (LLM) sowie dessen Compliance mit den Datenschutzgesetzen. Zu diesem Zweck testet der Prüfer das Modell mit gezielten Abfragen zu einem bestimmten Patienten und sorgfältig abgestimmten künstlich erzeugten Nichtpatienten. Durch die Beobachtung von Unterschieden in den Antwortmustern und Konfidenzwerten des Modells kann der Prüfer zuverlässig feststellen, ob die Akte dieses Patienten in den Trainingsdaten enthalten war, auch wenn keine tatsächlichen medizinischen Details preisgegeben werden.

Welche Art von Angriff wird hier simuliert?

A) Datenoffenlegung in der Modellausgabe
B) Indirekte Prompt-Injection
C) Membership Inference
D) Model Inversion (Modellinversion)

A) Falsch. Eine Datenoffenlegung in der Modellausgabe liegt vor, wenn das Modell sensible, in den Trainingsdaten oder im Runtime-Kontext enthaltene Informationen direkt preisgibt. In diesem Szenario werden keine tatsächlichen medizinischen Details offengelegt.
B) Falsch. Eine indirekte Prompt-Injection liegt vor, wenn ein Dritter adversariale Anweisungen in Inhalte einbettet, die abgerufen und in einen Prompt eingefügt werden, wodurch das Verhalten des Modells manipuliert wird. Das beschriebene Szenario enthält keine solche Injektion.
C) Richtig. Membership Inference liegt vor, wenn ein Angreifer dem Modell Eingaben vorlegt, die eine bestimmte Person identifizieren und anhand der Konfidenzsignale des Modells ermittelt, ob diese Person im Trainingsdatensatz vorkommt. (Literatur: A, Kapitel 2)
D) Falsch. Eine Model Inversion liegt vor, wenn ein Angreifer durch intensive Experimente einen Teil des Trainingsdatensatzes rekonstruiert, indem er Eingaben so optimiert, dass die Konfidenzsignale in der Modellausgabe maximiert werden. Bei dieser Art von Angriff werden sensible Attribute oder Datensätze aus dem Modell rekonstruiert, anstatt lediglich festzustellen, ob ein bestimmter Datensatz Teil der Trainingsdaten war.

12 / 40
Was zählt zu den **wesentlichen** Folgen eines erfolgreichen Angriffs auf die Model Exfiltration (Modellausleitung)?

A) Der Angreifer kann den ursprünglichen Trainingsdatensatz in Rohform extrahieren, der zur Erstellung des ursprünglichen Modells verwendet wurde und der möglicherweise sensible Datensätze enthält.
B) Der Angreifer erlangt direkte Kontrolle über das bereitgestellte Modell und kann dessen Gewichte und Parameter ohne vorherige Genehmigung ändern.
C) Der Angreifer erhält eine Kopie des Modells. Diese kann zur Entwicklung von Evasion Attacks (Umgehungsangriffe) genutzt werden, ohne die Abwehrmechanismen des Originals auszulösen.
D) Das ursprüngliche Modell steht legitimen Nutzern aufgrund von Resource Exhaustion (Ressourcenerschöpfung) nicht mehr zur Verfügung, was zu einer Verschlechterung der Dienstqualität und wiederholten Zeitüberschreitungen führt.

A) Falsch. Bei der Model Exfiltration wird das Verhalten des Modells anhand von Eingabe-Ausgabe-Paaren nachgebildet, der ursprüngliche Trainingsdatensatz wird jedoch nicht extrahiert. Das Extrahieren von Trainingsdaten stellt eine andere Bedrohung dar (Model Inversion (Modellinversion) oder Membership Inference).
B) Falsch. Die Model Exfiltration gewährt nicht die Möglichkeit, das ursprünglich bereitgestellte Modell zu modifizieren. Sie resultiert in einer Kopie der Funktionalität des Modells und nicht in der Kontrolle über das Original.
C) Richtig. Eine wesentliche Folge der Model Exfiltration besteht darin, dass der Angreifer die Replik des Modells für Evasion Attacks nutzen kann und dabei die im Originalsystem vorhandenen Mechanismen zur Ratenbegrenzung, Zugriffskontrolle und Erkennung umgeht. (Literatur: A, Kapitel 2)
D) Falsch. Ein Denial-of-Service (DoS) ist eine Folge der Resource Exhaustion, die eine eigenständige Bedrohung darstellt. Die Model Exfiltration hat keinen direkten Einfluss auf die Verfügbarkeit des Originalmodells.

13 / 40
Ein in der Cloud gehostetes Large Language Model (LLM) wird Opfer einer Sponge Attack (Sponge-Angriff). Angreifer senden sehr lange oder komplexe Eingaben, die zusätzliche Rechenleistung erfordern. Dies treibt die Kosten in die Höhe und verursacht Verzögerungen oder Ausfälle. Um dieser Situation entgegenzuwirken, möchte der KI-Sicherheitsingenieur zwei Kontrollmaßnahmen einführen:

- **Kontrollmaßnahme 1** validiert und bereinigt Eingaben, um böswillige Inhalte, wie beispielsweise ungewöhnlich große Eingaben, zurückzuweisen oder zu korrigieren.
- **Kontrollmaßnahme 2** begrenzt die pro Modelleingabe verbrauchten Rechenressourcen, um eine Überbeanspruchung zu verhindern.

Um welche beiden Kontrollmaßnahmen handelt es sich?

A) - Kontrollmaßnahme 1: Behandlung anomaler Eingaben
- Kontrollmaßnahme 2: Verschleierung von Konfidenzgrenzen

B) - Kontrollmaßnahme 1: Validierung von Eingaben im Hinblick auf Denial-of-Service (DoS)
- Kontrollmaßnahme 2: Ressourcenbegrenzung

C) - Kontrollmaßnahme 1: Zugriffskontrolle auf das Modell
- Kontrollmaßnahme 2: gegen Evasion Attacks (Umgehungsangriffe) robustes Modell

D) - Kontrollmaßnahme 1: Nutzungsüberwachung
- Kontrollmaßnahme 2: Ratenbegrenzung

A) Falsch. Die Behandlung anomaler Eingaben erkennt von der Verteilung der Trainingsdaten abweichende Eingaben und ist in erster Linie für Bedrohungen durch Evasion Attacks und Model Exfiltration (Modellausleitung) relevant. Die Verschleierung von Konfidenzgrenzen schränkt die Ausgabedaten ein, um die Fähigkeit des Angreifers zur Optimierung adversarialer Beispiele zu verringern.

B) Richtig. Dies sind die beiden in der Frage genannten Kontrollmaßnahmen, die speziell darauf ausgelegt sind, eine durch Eingaben verursachte Resource Exhaustion (Ressourcenerschöpfung) zu bekämpfen und die die allgemeinen Kontrollmaßnahmen ergänzen. (Literatur: A, Kapitel 2)

C) Falsch. Die Modellzugriffskontrolle schränkt ein, wer mit dem Modell interagieren darf und ist eine allgemeine Kontrollmaßnahme, die auf mehrere Eingabebedrohungen anwendbar ist. Ein gegen Evasion Attacks robustes Modell ist eine Kontrollmaßnahme während der Entwicklungsphase, die darauf ausgelegt ist, die Widerstandsfähigkeit des Modells gegen Evasion Attacks zu erhöhen.

D) Falsch. Bei der Nutzungsüberwachung geht es darum, die Modellnutzung, Eingaben, Ausgaben und das Systemverhalten zu beobachten, zu korrelieren und zu protokollieren, um Ereignisse oder Muster zu identifizieren, die auf einen (Cyber-)Sicherheitsvorfall hindeuten könnten. Bei der Ratenbegrenzung geht es darum, die Zugriffsrate (-häufigkeit) auf das Modell zu begrenzen.

14 / 40
Ein KI-Sicherheitsingenieur identifiziert die folgenden Bedrohungen für ein Unternehmen:

- **Adversariale Prompts** werden von Angreifern erstellt, um Fehlklassifizierungen zu verursachen.
- Eine **kompromittierte Erfassungspipeline** führt Datensätze mit Hintertüren in den Korpus ein, bevor das KI-Modell erstellt wird.
- Angreifer manipulieren **Labeling-Workflows** und ändern die Ground-Truth-Annotationen, um das Modellverhalten zu verzerren.
- **Trainingsdaten** werden ohne ordnungsgemäße Validierung aus externen oder Drittquellen bezogen.

Bei welcher Bedrohung besteht laut OWASP AI Exchange **kein** Risiko von Data Poisoning (Datenvergiftung) in der Entwicklungsphase?

A) Adversariale Eingaben
B) Kompromittierte Erfassungspipeline
C) Labeling-Workflows
D) Trainingsdaten

A) Richtig. Dies beschreibt adversariale Beispiele zum Zeitpunkt der Bereitstellung, nicht Data Poisoning in der Entwicklungsphase. (Literatur: A, Kapitel 3)
B) Falsch. Das Einschleusen manipulierter Elemente in den gesammelten Korpus vor der Modellerstellung führt zum Einbau einer Hintertür während der Entwicklung, was ein Risiko von Data Poisoning in der Entwicklungsphase darstellt.
C) Falsch. Die Manipulation von Labels während der Annotation vergiftet den Datensatz bereits vor der Modellerstellung, was einen Angriff während der Entwicklungsphase darstellt.
D) Falsch. Die Verwendung ungeprüfter externer Datensätze birgt das Risiko, dass vergiftete Beispiele in den Datenquellen enthalten sind. Die OWASP AI Exchange hebt Risiken in der Lieferkette und die Notwendigkeit einer Kontrolle der Datenqualität hervor.

15 / 40
Was ist laut OWASP AI Exchange ein Beispiel für **Direct Development-Time Model Poisoning** (direkte Modellvergiftung in der Entwicklungsphase)?

A) Angreifer erstellen während der Runtime adversariale Eingaben, um das Modell in die Irre zu führen
B) Angreifer manipulieren Gewichte und eine Trainingspipeline, um das Verhalten des Modells zu verändern.
C) Sensible Trainingsdaten werden unbeabsichtigt über die Modellausgaben offengelegt.
D) Ein plötzlicher Anstieg des Datenverkehrs überlastet den KI-Dienst, was zu Verlangsamungen oder Ausfallzeiten führt.

A) Falsch. Hierbei handelt es sich um einen adversarialen Angriff während der Runtime, nicht um eine Vergiftung während der Entwicklungsphase.
B) Richtig. Ein Direct Development-Time Model Poisoning beinhaltet die Kompromittierung der Entwicklungsumgebung, der Trainingspipeline oder der Modellparameter. Der Angreifer kann Logik einfügen, Gewichte modifizieren oder Trainingsverfahren verändern, um böswilliges Verhalten einzubetten. (Literatur: A, Kapitel 3)
C) Falsch. Dies bezieht sich auf Bedrohungen der Vertraulichkeit wie Model Inversion (Modellinversion) oder Datenoffenlegung, nicht auf Model Poisoning (Modellvergiftung).
D) Falsch. Dabei handelt es sich um ein Denial-of-Service (DoS)-Problem, nicht um Model Poisoning.

16 / 40
Was ist laut OWASP AI Exchange ein Beispiel für **Supply-Chain Model Poisoning** (Modellvergiftung in der Lieferkette)?

A) Angreifer kompromittieren Daten, Modelle oder Komponenten der Lieferkette vor deren Integration.
B) Angreifer manipulieren Eingaben während der Runtime, um das Modell dazu zu bringen, falsche Ergebnisse zu erzeugen.
C) Das Modell liefert aufgrund unzureichender Daten oder Trainingsqualität falsche Ergebnisse.
D) Das Modell gibt unbeabsichtigt sensible Trainingsdaten über seine Antworten preis.

A) Richtig. Der Angriff erfolgt vor der Bereitstellung des KI-Modells über Quellen von Dritten wie vortrainierte Modelle, externe Datensätze sowie Komponenten des maschinellen Lernens (ML) oder Pipelines. OWASP AI Exchange betont die Notwendigkeit eines Lieferkettenmanagements, um Vertrauen und Integrität zu gewährleisten. (Literatur: A, Kapitel 3)
B) Falsch. Dies beschreibt Evasion Attacks (Umgehungsangriffe) oder Prompt-Injection, nicht das Supply-Chain Model Poisoning. Diese Angriffe finden während der Nutzung des Modells statt, nicht während der Entwicklung oder Beschaffung.
C) Falsch. Falsche Ergebnisse aufgrund fehlender Daten oder mangelnder Trainingsqualität sind keine Angriffe. Beim Supply-Chain Model Poisoning handelt es sich ausdrücklich um vorsätzliche, böswillige Manipulationen.
D) Falsch. Dies bezieht sich auf Bedrohungen der Vertraulichkeit wie Model Inversion (Modellinversion) oder die Offenlegung sensibler Daten, nicht auf Model Poisoning (Modellvergiftung).

17 / 40
Ein KI-Entwicklungsteam führt eine Bewertung des Risikos seiner Entwicklungsumgebung durch. Es identifiziert zwei unterschiedliche Bedrohungen:

Bedrohung 1 betrifft den unbefugten Zugriff auf die zum Trainieren und Testen des KI-Modells verwendeten Datensätze.
Bedrohung 2 betrifft den unbefugten Zugriff auf die Modellparameter und -gewichte, die in einem kompromittierten Repository gespeichert sind.

Um welche Arten von Bedrohungen handelt es sich?

A) - Bedrohung 1: Development-Time Data Leak (Datenleck in der Entwicklungsphase)
- Bedrohung 2: Direct Development-Time Model Leak (direktes Modellleck in der Entwicklungsphase)

B) - Bedrohung 1: Direct Augmentation Data Leak (direktes Leck von Augmentierungsdaten)
- Bedrohung 2: Repository Leak (Repository-Leck)

C) - Bedrohung 1: Direct Runtime Data Leak (direktes Datenleck während der Runtime)
- Bedrohung 2: Direct Runtime Model Leak (direktes Modellleck während der Runtime)

D) - Bedrohung 1: Input Data Leak (Leck von Eingabedaten)
- Bedrohung 2: Source Code/Configuration Leak (Quellcode-Leck/Konfigurationsleck)

A) Richtig. Ein Development-Time Data Leak beinhaltet den unbefugten Zugriff auf Trainings- und Testdaten. Dadurch wird die Vertraulichkeit sensibler Daten oder des in diesen Datensätzen enthaltenen geistigen Eigentums verletzt. Ein Direct Development-Time Model Leak beinhaltet den unbefugten Zugriff auf Modellattribute (Parameter, Gewichte, Modellarchitektur), wodurch die Vertraulichkeit des geistigen Eigentums des Modells verletzt wird. (Literatur: A, Kapitel 3)

B) Falsch. Ein Direct Augmentation Data Leak entspricht nicht den Trainings- und Testdatensätzen und der Begriff Repository Leak ist zu vage. Es handelt sich konkret um ein Model Artifact Leak (Leck von Modellartefakten), nicht lediglich um ein generisches Repository-Problem.

C) Falsch. Die beschriebenen Bedrohungen sind keine Lecks während der Runtime. Sie treten in der Entwicklungsumgebung auf.

D) Falsch. Ein Input Data Leak bezieht sich im Allgemeinen auf das Durchsickern von benutzerbereitgestellten Eingaben, nicht auf Trainings- und Testdatensätze. Modellgewichte sind kein Quellcode/keine Konfiguration, sondern Modellartefakte.

18 / 40
Ein Sicherheitsteam überprüft die Runtime-Bedrohungslandschaft eines KI-Systems. Das KI-System ist in eine Webanwendung integriert. Das Team identifiziert zwei herkömmliche Sicherheitsbedrohungen:

- **Bedrohung 1:** Ein Angreifer kann Zugriff auf den Speicher des Produktionsservers erlangen und die Gewichte und Parameter des Modells extrahieren.
- **Bedrohung 2:** Ein Angreifer mit Zugriff auf das Modellregister ersetzt das bereitgestellte Modellartefakt durch eine mit einem Trojaner versehene Version.

Welche **KI-spezifischen** Folgen haben diese herkömmlichen Sicherheitsbedrohungen?

A) - Bedrohung 1 führt dazu, dass das System anhand schädlicher Daten automatisch neu trainiert wird.
- Bedrohung 2 zwingt das Modell dazu, gültige Prompts abzulehnen.

B) - Bedrohung 1 verändert die Modellarchitektur während der Runtime.
- Bedrohung 2 komprimiert die Gewichte unbemerkt auf eine geringere Präzision.

C) - Bedrohung 1 deaktiviert Sicherheitsfilter, wodurch das System von Grund auf toleranter wird.
- Bedrohung 2 macht die Ausgaben deterministisch und vorhersehbar.

D) - Bedrohung 1 ermöglicht Modelldiebstahl und potenzielle Rückschlüsse auf Trainingsdaten.
- Bedrohung 2 ermöglicht unentdeckte Manipulationen am Modell.

A) Falsch. Bei Bedrohung 1 geht es nicht um Vergiftung durch automatisches Umtrainieren, sondern um Modelldiebstahl und Dateninferenz. Bei Bedrohung 2 geht es um Integritätsmanipulation und Hintertüren, nicht um das Zurückweisen von Prompts.

B) Falsch. Änderungen an der Runtime-Architektur und eine stillschweigende Präzisionskomprimierung sind keine KI-spezifischen Folgen der Bedrohungen. Sie beschreiben eher unplausibles oder nicht böswilliges Verhalten statt Diebstahl oder verdeckte Manipulation.

C) Falsch. Die Deaktivierung von Sicherheitsfiltern entspricht einer Manipulation, nicht dem Exfiltrationsrisiko von Bedrohung 1. Die Determinierung von Ausgaben verkennt den Integritätscharakter von Bedrohung 2.

D) Richtig. Bedrohung 1 führt zu Modelldiebstahl und potenziellen Rückschlüssen auf Trainingsdaten. Bedrohung 2 führt zu verdeckten Integritätsangriffen (unentdeckte Modellmanipulation). (Literatur: A, Kapitel 4)

19 / 40
Ein Finanzdienstleistungsunternehmen setzt ein Modell für maschinelles Lernen (ML-Modell) ein, um betrügerische Transaktionen zu erkennen. Das ML-Modell wird in der Produktionsumgebung des Unternehmens gehostet und verarbeitet kontinuierlich Live-Transaktionsdaten. Es treten zwei Sicherheitsvorfälle auf:

- **Vorfall 1**: Ein Angreifer verschafft sich unbefugt Zugriff auf die Produktionsumgebung des Modells und verändert die Modellparameter so, dass bestimmte betrügerische Transaktionen nicht mehr als solche gekennzeichnet werden. Die Manipulation wird nicht erkannt und der Angreifer nutzt dies zu seinem Vorteil.
- **Vorfall 2**: Ein Angreifer nutzt eine Sicherheitslücke im Produktionssystem und kann so das Modell replizieren und dessen Verhalten für weitere Angriffe analysieren.

Welche Arten von Sicherheitsvorfällen nach OWASP AI Exchange werden hier beschrieben?

A) - Vorfall 1: Data Poisoning (Datenvergiftung)
- Vorfall 2: Direct Runtime Model Poisoning (direkte Modellvergiftung während der Runtime)

B) - Vorfall 1: Data Poisoning
- Vorfall 2: Evasion Attack (Umgehungsangriff)

C) - Vorfall 1: Direct Runtime Model Poisoning
- Vorfall 2: Direct Runtime Model Leak (direktes Modellleck während der Runtime)

D) - Vorfall 1: Direct Runtime Model Leak
- Vorfall 2: Evasion Attack

A) Falsch. Ein Direct Runtime Model Leak bezeichnet den Sicherheitsvorfall 2 korrekt. Der Angreifer extrahiert Modellparameter und -architektur und gefährdet damit die Vertraulichkeit des Modells. Dies ermöglicht die Nachbildung und weitere Ausnutzung. Data Poisoning tritt jedoch während der Entwicklungsphase durch Manipulation der Trainingsdaten auf. Das Szenario beschreibt eine Kompromittierung während der Runtime, nicht eine Manipulation in der Trainingsphase.

B) Falsch. Data Poisoning tritt während der Entwicklungsphase durch Manipulation der Trainingsdaten auf. Das Szenario beschreibt eine Kompromittierung während der Runtime, nicht eine Manipulation in der Trainingsphase. Eine Evasion Attack beinhaltet die Manipulation von Eingaben, nicht die Änderung des Modells selbst. Im Szenario wird das Modell direkt verändert.

C) Richtig. „Direct Runtime Model Poisoning“ identifiziert den Sicherheitsvorfall 1 korrekt. Der Angreifer modifiziert das bereitgestellte Modell und verändert dessen Verhalten. Hierbei handelt es sich um einen Angriff auf die Modellintegrität, der als „Runtime Poisoning“ (Neuprogrammierung) bezeichnet wird. „Direct Runtime Model Leak“ identifiziert den Sicherheitsvorfall 2 korrekt. Der Angreifer extrahiert Modellparameter und Modellarchitektur und gefährdet damit die Vertraulichkeit des Modells. (Literatur: A, Kapitel 4)

D) Falsch. „Direct Runtime Model Leak“ identifiziert den Sicherheitsvorfall 2 korrekt. Der Angreifer extrahiert Modellparameter und -architektur und gefährdet damit die Vertraulichkeit des Modells. Dies ermöglicht die Nachbildung und weitere Ausnutzung. Eine Evasion Attack beinhaltet jedoch die Manipulation von Eingaben, nicht die Änderung des Modells selbst. Im Szenario wird das Modell direkt verändert. (Literatur: A, Kapitel 4)

20 / 40
Eine Einrichtung des Gesundheitswesens setzt einen KI-Assistenten ein, der dem medizinischen Fachpersonal dabei hilft, Patientenakten zusammenzufassen und Empfehlungen auszusprechen. Das System ist mit internen Datenbanken und externen Tools vernetzt und das medizinische Fachpersonal stützt sich im täglichen Arbeitsablauf auf die von ihm generierten Ergebnisse. Es werden zwei Bedrohungen beobachtet:

- **Bedrohung 1**: Der KI-Assistent generiert eine Antwort, die ein verstecktes, bösartiges Skript enthält, das in HTML-Inhalte eingebettet ist. Bei Anzeige der Ausgabe wird das Skript ausgeführt und sendet Session-Daten an den Angreifer.
- **Bedrohung 2**: Ein Angreifer nutzt schwache Zugriffskontrollen im System aus, um sensible Patientendaten abzufangen und abzurufen.

Welche Bedrohungsarten nach OWASP AI Exchange werden hier beschrieben?

A) - Bedrohung 1: Data Poisoning (Datenvergiftung)
- Bedrohung 2: Input Data Leak (Leck von Eingabedaten)

B) - Bedrohung 1: Data Poisoning
- Bedrohung 2: Model Inversion (Modellinversion)

C) - Bedrohung 1: Ausgabe mit klassischer Injection
- Bedrohung 2: Leck von Eingabedaten

D) - Bedrohung 1: Ausgabe mit klassischer Injection
- Bedrohung 2: Model Inversion

A) Falsch. Bedrohung 2 wird korrekt als Input Data Leak identifiziert. Dem Modell bereitgestellte sensible Daten werden offengelegt oder abgefangen, wodurch die Vertraulichkeit verletzt wird. Data Poisoning tritt jedoch während der Entwicklungsphase durch die Veränderung von Trainingsdaten auf. Das Szenario beschreibt eine klassische Injection, nicht die Manipulation des Trainings.

B) Falsch. Data Poisoning tritt während der Entwicklungsphase durch die Manipulation von Trainingsdaten auf. Das Szenario beschreibt eine klassische Injection, nicht eine Manipulation des Trainings. Bei der Model Inversion werden Trainingsdaten aus den Ausgaben abgeleitet. Das Szenario beschreibt keine Situation, in der eine Modellinferenz stattgefunden hat.

C) Richtig. Bedrohung 1 wird korrekt als Ausgabe mit klassischer Injection identifiziert. Die KI-Ausgabe enthält bösartigen ausführbaren Inhalt, in diesem Fall eine Skript-Injection. Bedrohung 2 wird korrekt als Input Data Leak identifiziert. Sensible, dem Modell bereitgestellte Daten werden offengelegt oder abgefangen, wodurch deren Vertraulichkeit verletzt wird. (Literatur: A, Kapitel 4)

D) Falsch. Bedrohung 1 wird korrekt als Ausgabe mit klassischer Injection identifiziert. Die KI-Ausgabe enthält bösartigen ausführbaren Inhalt, in diesem Fall eine Skript-Injection. Bei der Model Inversion werden jedoch Trainingsdaten aus den Ausgaben abgeleitet. Das Szenario beschreibt keine Situation, in der eine Modellinferenz stattgefunden hat.

21 / 40
Ein Unternehmen nutzt ein RAG-System (Retrieval-Augmented Generation), um Fragen von Mitarbeitenden zu internen Richtlinien zu beantworten. Das KI-System ruft Dokumente aus einer Vektordatenbank ab und fügt diese dem Prompt hinzu, bevor es sie an das KI-Modell übermittelt.

Bei einer Sicherheitsüberprüfung wird festgestellt, dass sich ein Angreifer unbefugten Zugriff auf die Vektordatenbank verschafft hat. Der Angreifer hat die gespeicherten Augmentierungsdokumente eingesehen und ein legitimes Richtliniendokument durch ein anderes ersetzt, das falsche Informationen enthält. Mitarbeitende, die das KI-System zu dieser Richtlinie befragen, erhalten nun falsche Hinweise.

Welche Sicherheitsbedrohungen veranschaulicht dieses Szenario?

A) - Bedrohung 1: Augmentation Data Leak (Leck von Augmentierungsdaten)
- Bedrohung 2: Ausgabe mit klassischer Injection

B) - Bedrohung 1: Manipulation von Augmentierungsdaten
- Bedrohung 2: Augmentation Data Leak

C) - Bedrohung 1: Membership Inference
- Bedrohung 2: Manipulation von Augmentierungsdaten

D) - Bedrohung 1: Ausgabe mit klassischer Injection
- Bedrohung 2: Membership Inference

A) Falsch. Eine Ausgabe mit klassischer Injection beinhaltet bösartigen ausführbaren Inhalt, der in die Modellausgabe eingebettet ist. Dies ist in diesem Szenario nicht der Fall. Zwar wurde das Augmentation Data Leak korrekt identifiziert, aber die Gesamtantwort ist nicht korrekt.

B) Richtig. Eine Manipulation von Augmentierungsdaten tritt auf, wenn ein Angreifer die in einer Vektordatenbank gespeicherten Augmentierungsdaten modifiziert und falsche Informationen einfügt, die das Verhalten des Modells verändern. Ein Augmentation Data Leak entsteht, weil sich ein Angreifer unbefugt Zugriff auf die in der Vektordatenbank gespeicherten Augmentierungsdaten verschafft hat. Die Vektordatenbank stellt einen zusätzlichen Speicherort dar, an dem auf Augmentierungsdaten zugegriffen oder diese verändert werden können, wodurch eine neue Angriffsfläche entsteht und die Risiken über das ursprüngliche Dokumenten-Repository hinaus erweitert werden. (Literatur: A, Kapitel 4)

C) Falsch. Membership Inference ist ein Angriff, bei dem ein Angreifer durch die Analyse von Konfidenzsignalen ermittelt, ob bestimmte Daten im Trainingssatz vorkommen. Dies wird im Szenario nicht beschrieben. Zwar wird die Manipulation von Augmentierungsdaten korrekt identifiziert, aber die Zuordnung zur Membership Inference ist falsch.

D) Falsch. Eine Ausgabe mit klassischer Injection bezieht sich auf bösartigen Code, der in die Modellausgabe eingebettet ist. Membership Inference ist ein Angriff, bei dem ein Angreifer durch die Analyse von Konfidenzsignalen ermittelt, ob bestimmte Daten im Trainingssatz vorkommen. Keine dieser Sicherheitsbedrohungen ist in dem Szenario beschrieben.

22 / 40
Allgemeine Governance-Kontrollen sollten implementiert werden, um eine wirksame Sicherheitsaufsicht über KI-Systeme zu gewährleisten.

Welcher Ansatz verdeutlicht dies **am besten**?

A) Die Festlegung klarer Richtlinien, Rollen und Risikomanagementpraktiken als Leitlinien für die sichere Entwicklung, Bereitstellung und Überwachung von KI-Systemen
B) Die Einbindung von Sicherheitskontrollen in die KI-Systeme, um die Security-by-Design-Prinzipien bereits in der Entwurfsphase des KI-Systems anzuwenden
C) Die Einhaltung von Mindestsicherheitsanforderungen, da KI-Systeme standardmäßig sicherer sind und nicht mehrere Sicherheitsebenen erfordern
D) Die Trennung der KI-Governance-Kontrollen von den Informationssicherheitskontrollen, um sicherzustellen, dass diese beiden unterschiedlichen Konzepte gleichermaßen abgedeckt sind

A) Richtig. Bei Governance-Kontrollen geht es um die Festlegung und Durchsetzung von Richtlinien, Rollen, Verantwortlichkeiten und Risikomanagement innerhalb eines Informationssicherheits-Managementsystems (ISMS), um die durchgängige Aufsicht über den gesamten KI-Lebenszyklus hinweg zu gewährleisten. (Literatur: A, Kapitel 1)
B) Falsch. Die Anwendung der Security-by-Design-Prinzipien in der Entwurfsphase von KI-Systemen ist ein wichtiger Bestandteil der allgemeinen Governance-Kontrollen, reicht jedoch nicht aus. Governance umfasst Richtlinien, Verantwortlichkeiten, Risiken, Überwachung und kontinuierliche Verbesserung.
C) Falsch. KI-Systeme sind nicht standardmäßig sicher. Sich auf Mindestanforderungen zu verlassen, lässt den für sich weiterentwickelnde KI-Risiken erforderlichen mehrschichtigen Schutz und die kontinuierliche Aufsicht außer Acht.
D) Falsch. Die Trennung der KI-Governance von der Informationssicherheit führt zu Silos und Lücken. Eine allgemeine Best Practice ist die Einbindung in Frameworks für das Unternehmensrisikomanagement und für Informationssicherheits-Managementsysteme (ISMS), um kohärente Kontrollmaßnahmen und Verantwortlichkeit zu gewährleisten.

23 / 40
Eine mittelgroße Marketingagentur verfügt über wenig KI-Erfahrung und hat noch keine formelle KI-Governance etabliert. Sie möchte ein generatives KI-Tool zum Verfassen von Kunden-E-Mails einsetzen. Das KI-Tool benötigt Zugriff auf Kundennamen und E-Mail-Adressen. Das Unternehmen sollte vor Nutzung des KI-Tools eine KI-Governance und eine Aufsicht über die KI-Sicherheit einrichten.

Was ist das **absolute Minimum**, das das Unternehmen für die Aufsicht über die KI-Sicherheit unternehmen sollte?

A) Red-Team-Tests an dem E-Mail-System der generativen KI durchführen, um Sicherheitsrisiken vor der Einführung des Systems zu mindern
B) Einen KI-Governance-Spezialisten einstellen, der eine Bedrohungsanalyse durchführt und Richtlinien und Kontrollmaßnahmen für die Einführung des Systems festlegt
C) Eine Bestandsaufnahme der geplanten KI-Nutzung erstellen und eine Risikoanalyse durchführen, um Bedrohungen, Kontrollmaßnahmen und Verantwortlichkeiten zu identifizieren
D) Rechtsberatung von einem Unternehmen einholen, das sich auf KI und die damit verbundenen Anforderungen der KI-Verordnung spezialisiert hat
E) Mit einem Pilotprojekt starten und das KI-E-Mail-System hinsichtlich Leistung, Missbrauch und Vorfällen anhand von Protokollen überwachen

A) Falsch. Red-Team-Tests stellen nicht das absolute Minimum für die Aufsicht über die KI-Sicherheit dar. Sie sind eine nachgelagerte Maßnahme zur Absicherung, die auf einem klar definierten Umfang, grundlegenden Kontrollen und eindeutigen Zuständigkeiten aufbaut, die im Rahmen der anfänglichen Governance festgelegt wurden.
B) Falsch. Die Einstellung eines Spezialisten für KI-Governance ist eine Personalentscheidung, kein Mindestschritt der Governance.
C) Richtig. Diese Option stellt das absolute Minimum dar, das im Rahmen der Governance unternommen werden sollte, um vor der Bereitstellung des Systems die Risiken, Bedrohungen, Kontrollmaßnahmen und Verantwortlichkeiten zu identifizieren und so die Aufsicht über die KI-Sicherheit zu gewährleisten. (Literatur: A, Kapitel 1)
D) Falsch. Eine rechtliche Prüfung ist wichtig, gehört jedoch zum Bereich der Compliance und kann die Governance der KI-Sicherheit nicht ersetzen. Eine Verzögerung der Governance führt dazu, dass Risiken im Zusammenhang mit personenbezogenen Daten und deren Missbrauch unkontrolliert bleiben.
E) Falsch. Pilotprojekte und Überwachung sind operative und reaktive Maßnahmen. Die Durchführung eines Pilotprojekts an einem kleinen Kundensegment ist eine Implementierungsstrategie. Ohne vorherige Governance fehlen bei Pilotprojekten und Überwachung definierte Leitplanken, Rollen und Kontrollmaßnahmen zum Schutz der CRM-Daten.

24 / 40
Wie sollte der Geltungsbereich allgemeiner Governance-Kontrollmaßnahmen im Bereich der KI-Sicherheit **am besten** verstanden werden?

A) Als eine Sammlung von Verschlüsselungsmethoden, um die Sicherheit von KI-Ergebnissen nach deren Generierung zu gewährleisten
B) Als eine Reihe übergreifender Kontrollmaßnahmen, die für alle KI-Bedrohungen und Lebenszyklusphasen gelten
C) Als auf die Trainingsphase des Modells beschränkte Kontrollmaßnahmen, die sich auf die Qualität der Datensätze und Optimierungstechniken konzentrieren
D) Als optionale Maßnahmen, die Organisationen ergreifen können, wenn die KI-Systeme keine sensiblen Daten verarbeiten
E) Als technische Schutzmaßnahmen, die ausschließlich während der Runtime angewendet werden, um bösartige Eingaben zu erkennen

A) Falsch. Verschlüsselung ist lediglich eine mögliche technische Kontrollmaßnahme und stellt keine umfassende Governance dar. Allgemeine Kontrollmaßnahmen beschränken sich nicht auf Kryptografie oder Ausgaben.
B) Richtig. Governance ist übergreifend und bezieht sich auf Richtlinien, Rollen, Verantwortlichkeit und Risikomanagement, die für alle Bedrohungen und alle Lebenszyklusphasen gelten. (Literatur: A, Kapitel 1)
C) Falsch. Governance gilt für den gesamten Lebenszyklus, nicht nur für die Trainingsphase.
D) Falsch. Governance-Kontrollen sind weder optional noch auf die Verarbeitung sensibler Daten beschränkt.
E) Falsch. Allgemeine Kontrollmaßnahmen beschränken sich nicht auf die Runtime oder technische Abwehrmaßnahmen. Sie erstrecken sich über organisatorische, verfahrensbezogene und Governance-Schichten.

25 / 40
Ein Unternehmen im Gesundheitswesen beschließt, für eine Anwendung ein von einem Drittanbieter bereitgestelltes, vorgefertigtes KI-Modell einzusetzen. Die Anwendung verarbeitet Eingaben der Nutzer und ruft externe Daten ab.

Der KI-Sicherheitsingenieur hat Bedenken hinsichtlich der Sicherheit des Modells. Die Aufteilung der Sicherheitsverantwortlichkeiten zwischen dem Anbieter des KI-Modells und dem Unternehmen im Gesundheitswesen wurde nicht in einer Sicherheitsstrategie festgelegt.

Was ist in Bezug auf diese Verantwortlichkeiten **richtig**?

A) Werden vorgefertigte KI-Modelle verwendet, können die meisten Sicherheitskontrollen an den Anbieter delegiert werden, da das Training und das Hosting durch den Anbieter die Sicherheitsrisiken abdecken.
B) Der Betreiber ist für alle Sicherheitskontrollen verantwortlich, wenn er das KI-Modell in seine Systeme integriert und die Sicherheit vor der Bereitstellung berücksichtigt.
C) Der Anbieter hat das KI-Modell trainiert und verwaltet daher die Kontrollmaßnahmen auf Modellebene, während der Betreiber die Kontrollmaßnahmen auf Anwendungsebene verwalten sollte.
D) Der Anbieter ist für alle Sicherheitskontrollen verantwortlich, da er das KI-Modell trainiert hat und es hostet; der Anbieter sollte daher für den Umgang mit allen Risiken und Angriffen verantwortlich sein.

A) Falsch. Prompt-Injection und Leckage der Daten im Zusammenhang mit externen Tools oder der Retrieval-Augmented Generation (RAG) sind in erster Linie Risiken auf Anwendungsebene und können nicht vollständig an den Anbieter delegiert werden. Hosting und Training allein reichen nicht aus, um diese Risiken zu beherrschen.
B) Falsch. Der Anbieter trainiert und optimiert das Modell und definiert grundlegende Safety-Kontrollen. Er trägt einen Teil der Sicherheitsverantwortung, sodass diese zwischen den beteiligten Parteien geteilt ist.
C) Richtig. Der Anbieter trainiert und optimiert das Modell und ist für die Kontrollmaßnahmen auf Modellebene verantwortlich. Der Betreiber muss sich bei der Verwendung von vorgefertigten Modellen dennoch der Bedrohungen bewusst sein und ist für die Kontrollmaßnahmen auf Anwendungsebene verantwortlich. (Literatur: A, Kapitel 0)
D) Falsch. Obwohl der Anbieter das Modell trainiert und hostet, ist er nicht für alle Sicherheitskontrollen verantwortlich. Er trägt einen Teil der Sicherheitsverantwortung, sodass diese zwischen den beteiligten Parteien geteilt ist.

26 / 40
Eine Organisation hat eine kundenorientierte KI-Anwendung bereitgestellt, die ein Large Language Model (LLM) eines Drittanbieters (Third-Party Provider) nutzt. Der Zugriff auf das Modell erfolgt über einen verwalteten API-Dienst (Application Programming Interface). Das Modell wurde von der Organisation nicht weiter trainiert oder modifiziert.

Das Sicherheitsteam stellt fest, dass konstruierte Nutzereingaben dazu führen können, dass das Modell seine konfigurierten Verhaltensgrenzen ignoriert und Ausgaben erzeugt, die gegen die Inhaltsbeschränkungen der Organisation verstoßen. Das Team überlegt, welche Kontrolle implementiert werden sollte, die in die Zuständigkeit der Organisation als Betreiber (Deployer) fällt und nicht in die des Drittanbieters.

Welche Kontrolle sollte das Sicherheitsteam implementieren, um das Problem zu beheben?

A) Eine Ausgabevalidierungsschicht (Output Validation Layer) einfügen, die alle Modellantworten überprüft, bevor diese die Endnutzer erreichen, und alle Ausgaben blockiert, die gegen die Inhaltsrichtlinie der Organisation verstoßen
B) Die Inferenz-Pipeline des Anbieters ändern, um strengere Generierungsbeschränkungen durchzusetzen, die die Erzeugung von Antworten einschränken und das Entstehen unzulässiger Inhalte verhindern
C) Den Anbieter auffordern, das bereitgestellte Modell mithilfe zusätzlicher Daten zur Ablehnung unerlaubter Anfragen und zur Richtlinienkonformität neu zu trainieren und sicherheitsseitig abzustimmen, um seine Widerstandsfähigkeit gegen Jailbreak-Eingaben zu verbessern
D) Vom Anbieter verlangen, dass dieser die Regeln zur Inhaltsmoderation und deren Durchsetzung für den gesamten Dienst überarbeitet und verschärft, damit unsichere Ausgaben vor der Auslieferung global gefiltert werden

A) Richtig. Betreiber sind für die Kontrollen verantwortlich, die um die KI-Anwendung und deren Ausgaben herum implementiert werden, während Drittanbieter das zugrunde liegende Modelltraining und das Inferenzverhalten verwalten. Eine Ausgabevalidierungsschicht (Output Validation Layer) ist eine Kontrolle auf Seiten des Betreibers, die richtlinienwidrige Antworten blockieren kann, bevor sie die Endnutzer erreichen. (Literatur: A, Kapitel 0)
B) Falsch. Die Änderung der Inferenz-Pipeline des Anbieters liegt nicht im Einflussbereich des Betreibers, wenn eine von einem Drittanbieter verwaltete API genutzt wird. Er kann bestenfalls die zugänglichen Parameter anpassen, nicht jedoch die Pipeline selbst ändern.
C) Falsch. Das erneute Trainieren oder die Feinabstimmung des Modells für die Sicherheit liegt in der Verantwortung des Anbieters. Der Betreiber kann dies zwar anfordern, bei einem verwalteten API-Modell jedoch nicht selbst implementieren.
D) Falsch. Globale Moderationsrichtlinien und anbieterweite Filterkontrollen werden vom Drittanbieter verwaltet, nicht von der Betreiberorganisation.

27 / 40
Ein Team bereitet Trainingsdaten für einen generativen KI-Assistenten vor. Es stellt fest, dass es nicht alle Felder und Daten benötigt, die derzeit für die Leistung des KI-Modells verwendet werden. Als nächsten Schritt erwägt das Team Kontrollen zur Begrenzung sensibler Daten sowie zur Verbesserung der Vertraulichkeit und Integrität.

Wie lassen sich Vertraulichkeit und Integrität **am besten** verbessern?

A) Alle Felder in den Trainingsdaten beibehalten und die Daten im Ruhezustand sowie in Backups verschlüsseln
B) Identifikatoren in den Trainingsdaten beibehalten, um künftige Aufgaben und Prozesse des Neutrainings zu vereinfachen
C) Felder und Daten entfernen, die für das Modelltraining und die Modellleistung nicht erforderlich sind
D) Ausschließlich synthetische Daten für alle Trainings- und Evaluierungsphasen generativer KI verwenden

A) Falsch. Die Verschlüsselung im Ruhezustand schützt gespeicherte Daten, verringert aber nicht die unnötige Offenlegung sensibler Daten (Sensitive Data Exposure) und minimiert auch nicht die Angriffsfläche. Sie verbessert die Vertraulichkeit nur im Ruhezustand und trägt kaum zur Integrität bei.
B) Falsch. Das Beibehalten von Identifikatoren in den Trainingsdaten erhöht das Offenlegungsrisiko (Exposure Risk). Es untergräbt die Vertraulichkeit und kann die Integrität beeinträchtigen.
C) Richtig. Die Datenminimierung entfernt unnötige Felder und Datensätze und verringert so die Angriffsfläche der Daten. Sie stärkt die Integrität, indem sie Angreifern weniger Möglichkeiten zur Manipulation der Daten bietet und verhindert, dass sich das Modell auf irrelevante Muster stützt. (Literatur: A, Kapitel 1)
D) Falsch. Die Verwendung synthetischer Daten ist nicht immer machbar oder erforderlich und garantiert keine Verbesserung von Vertraulichkeit und Integrität. Synthetische Daten können weiterhin sensible Muster enthalten oder andere Risiken mit sich bringen.

28 / 40
Ein Produktteam trainiert ein KI-Modell für den Kundensupport. Der aktuelle Trainingsdatensatz enthält vollständige Kundenadressen, Telefonnummern und Alteinträge, die für das Erreichen der Leistungsziele nicht benötigt werden, künftig aber nützlich sein könnten. Nachdem eine Red-Team-Übung das Risiko von Datenlecks und Manipulation durch memorisierte Details aufgezeigt hat, beschließt das Team, die Offenlegung (Exposure) zu verringern. Dazu entfernt es unnötige Felder und ganze Einträge, sodass der Datensatz keine sensiblen Informationen mehr enthält.

Welche **allgemeine** Kontrolle wurde in dieser Situation zur Begrenzung sensibler Daten implementiert?

A) Zulässige Daten
B) Datenminimierung
C) Trainingsdaten verschleiern
D) Kurze Aufbewahrung

A) Falsch. Zulässige Daten definieren, welche Kategorien erlaubt sind, doch das Szenario erfordert das Löschen unnötiger Datensätze, was der Datenminimierung entspricht.
B) Richtig. Bei der Datenminimierung geht es darum, nicht benötigte Einträge aus dem Trainingsdatensatz zu entfernen, sodass keine sensiblen Daten vorhanden sind und somit auch nicht durchsickern können. (Literatur: A, Kapitel 1)
C) Falsch. Das Verschleiern von Trainingsdaten transformiert die Daten, bewahrt sie jedoch weiterhin auf. Daher können im Gegensatz zur Entfernung weiterhin Risiken bestehen bleiben.
D) Falsch. Die kurze Aufbewahrung begrenzt, wie lange Daten aufbewahrt werden, nicht jedoch, welche Einträge im Trainingsdatensatz enthalten sind.

29 / 40
Ein Einzelhandelsunternehmen beschließt, die Zahlungsdaten seiner Kunden nicht mehr zu speichern, sobald eine Transaktion abgeschlossen ist.

Inwiefern **verringert** diese Maßnahme das Risiko einer Datenoffenlegung (Data Exposure)?

A) Sie macht es unnötig, Daten während der Übertragung oder bei ihrer Archivierung zum Zwecke der Auditierbarkeit zu verschlüsseln
B) Sie gewährleistet die Compliance mit allen relevanten Datenschutzvorschriften und verringert die Komplexität
C) Sie gewährleistet, dass unbefugte Nutzer nicht auf das Unternehmensnetzwerk und die Datenbank zugreifen können
D) Sie verhindert, dass im System vorgehaltene Daten durchsickern, rekonstruiert oder abgeleitet werden können

A) Falsch. Die Datenminimierung macht eine Verschlüsselung nicht überflüssig. Übertragene Daten müssen weiterhin geschützt werden, unabhängig davon, wie viele Daten gespeichert werden.
B) Falsch. Die Datenminimierung unterstützt Bemühungen um Compliance, gewährleistet jedoch nicht automatisch die vollständige Compliance mit allen Datenschutzvorschriften, die zahlreiche zusätzliche Anforderungen beinhalten.
C) Falsch. Die Begrenzung gespeicherter Daten garantiert nicht, dass unbefugte Nutzer nicht auf das Netzwerk oder die Datenbank zugreifen können. Die Netzwerkzugangskontrolle ist eine separate Sicherheitsmaßnahme, die in keinem Zusammenhang mit der Datenminimierung steht.
D) Richtig. Wenn Daten nicht erhoben oder aufbewahrt werden, können sie nicht aus dem System offengelegt, rekonstruiert oder abgeleitet werden. Dies verringert unmittelbar die Auswirkungen von Datensatzdiebstahl oder unbefugtem Zugriff. (Literatur: A, Kapitel 1)

30 / 40
Ein regionales Krankenhausnetzwerk nutzt einen KI-Agenten, um eingehende E-Mails von Lieferanten zu lesen, Antworten zu formulieren und diese automatisch zu versenden. Bei den nach einem kürzlich aufgetretenen Phishing-Sicherheitsvorfall durchgeführten Sicherheitstests stellten die Tester fest, dass eine Prompt-Injection den KI-Agenten dazu bringen könnte, seine Regeln zu ignorieren und unbefugte Antworten zu versenden.

Welche Entwurfsentscheidung eignet sich **am besten**, um die Auswirkungen dieses unerwünschten Verhaltens zu begrenzen?

A) Alle Integrationen von KI-Agenten deaktivieren und die meisten Prozesse wieder auf vollständig manuelle Arbeitsabläufe umstellen
B) Die Schulungen der Nutzer zur Prompt-Sicherheit erweitern und die Notwendigkeit der Aufsicht beim agentengestützten E-Mail-Versand betonen
C) Weitreichende autonome Versandberechtigungen beibehalten und auf ein fortschrittlicheres generatives Modell upgraden
D) Die Berechtigungen des KI-Agenten auf aufgabenspezifische Bereiche beschränken und für Hochrisiko-Aktionen eine Genehmigung fordern

A) Falsch. Dies beseitigt die Automatisierung, anstatt Kontrollen zu implementieren, um die Funktionalität sicher beizubehalten.
B) Falsch. Das Bewusstsein der Nutzer hilft zwar, setzt jedoch keine KI-Runtime-Kontrollen durch.
C) Falsch. Eine Verbesserung der Modellqualität verringert den Auswirkungsradius nicht ausreichend, wenn das generative KI-System weiterhin über weitreichende Berechtigungen verfügt.
D) Richtig. Es werden Kontrollen nach dem Prinzip des Least Privilege und Human-in-the-Loop-Kontrollen empfohlen, um die Auswirkungen bei einer Prompt-Injection zu begrenzen. Das aufgabenspezifische Prinzip des Least Privilege schränkt die Handlungsmöglichkeiten des Agenten ein, während die Genehmigung von Hochrisiko-Aktionen für zusätzliche Aufsicht sorgt und unbefugte Antworten verhindert, selbst wenn das Modell getäuscht wird. (Literatur: A, Kapitel 1)

31 / 40
Wie lassen sich die Auswirkungen unerwünschten Modellverhaltens **am besten** auf das absolute Minimum beschränken?

A) Indem man Prompt-Hardening, rollenbasierten Zugriff und geplante Audits der Modellausgaben hervorhebt
B) Indem man sich auf automatisierte Ausgabefilter sowie die Protokollierung von Ausnahmen verlässt und gelegentliche Stichprobenprüfungen bezüglich Fairness durchführt
C) Indem man auf unerwünschten Bias testet und Kontrollen wie Aufsicht, Least Privilege und kontinuierliche Validierung anwendet
D) Indem man eine strenge Eingabevalidierung, eingeschränkte Tool-Berechtigungen und regelmäßige Human-in-the-Loop-Überprüfungen nutzt

A) Falsch. Diese Maßnahmen sind zwar nützlich, aber unvollständig. Prompt-Hardening ist anfällig, die rollenbasierte Zugriffskontrolle (RBAC) ohne ein umfassenderes Programm nach dem Prinzip des Least Privilege ist begrenzt, und geplante Audits finden nur periodisch statt und nicht kontinuierlich, was Lücken hinterlässt und systematische Tests auf Bias vermissen lässt.
B) Falsch. Automatisierte Ausgabefilter können umgangen werden, die Protokollierung von Ausnahmen verhindert keinen Schaden, und gelegentliche Stichproben zur Überprüfung der Fairness sind im Vergleich zur kontinuierlichen Validierung und zu strukturierten Tests auf Bias unzureichend.
C) Richtig. Kontrollen zur Begrenzung der Auswirkungen unerwünschten Modellverhaltens sind: Aufsicht, Least Model Privilege, Modell-Alignment, KI-Transparenz, Erklärbarkeit, kontinuierliche Validierung und Tests auf unerwünschten Bias. (Literatur: A, Kapitel 1)
D) Falsch. Eine strenge Eingabevalidierung, eingeschränkte Tool-Berechtigungen und regelmäßige Überprüfungen durch Menschen sind zwar hilfreich, doch in dieser Aufzählung fehlen die kontinuierliche Validierung und gezielte Tests auf Bias. Regelmäßige Überprüfungen lassen Lücken offen und sorgen nicht für eine umfassende, fortlaufende Reduzierung des Risikos.

32 / 40
Die Begrenzung unerwünschten Modellverhaltens ist eine entscheidende Strategie zur Begrenzung der Auswirkungen und zur Kontrolle des Auswirkungsradius.

Wie **verbessert** diese Strategie die Leistung und **verringert** das Risiko?

A) Sie konzentriert sich auf Angriffe, die KI-Systeme beeinträchtigen können, und beschränkt die Aufsicht von Systemleistungsfaktoren und Genauigkeit
B) Sie implementiert Kontrollen, die kontinuierliche Tests und Validierungen mit der Zeit überflüssig machen und so die Komplexität reduzieren
C) Sie verbessert die betriebliche Zuverlässigkeit und die Ressourceneffizienz und minimiert gleichzeitig Sicherheitsvorfälle und unnötigen Rechenaufwand durch Anfragen
D) Sie erhöht die Modellautonomie, sodass das Modell mehr Ausgaben generieren kann, um die Systemleistung kontinuierlich zu verbessern

A) Falsch. KI-Transparenz ist eine der Kontrollen zur Begrenzung unerwünschten Verhaltens. Sie steht in direktem Zusammenhang mit der Genauigkeit, da Angriffe oder Drifts, die unerwünschtes Verhalten verursachen, Probleme der Genauigkeit darstellen.
B) Falsch. Regelmäßige Tests und Leistungsvalidierungen stellen sicher, dass das Modell im Laufe der Zeit mit seinem vorgesehenen Verwendungszweck im Einklang bleibt, und begrenzen die Auswirkungen unerwünschten Verhaltens. Das Testen auf Bias ist eine der Kontrollen zur Begrenzung der Auswirkungen unerwünschten Verhaltens in KI-Systemen. Kontinuierliche Tests und Validierungen im Laufe der Zeit werden dringend empfohlen.
C) Richtig. Die Begrenzung unerwünschten Verhaltens reduziert schädliche Ausgaben und Sicherheitsvorfälle, verbessert die Zuverlässigkeit und steigert die Ressourceneffizienz, indem Wiederholungsversuche, unnötige Tool-Aufrufe und verschwendete Rechenressourcen reduziert werden. (Literatur: A, Kapitel 1)
D) Falsch. Eine Erhöhung der Modellautonomie führt nicht zu einer Leistungssteigerung. Der Autonomiegrad des Modells sollte unter Kontrolle gehalten werden, damit unerwünschtes Modellverhalten begrenzt werden kann.

33 / 40
Was ist im größeren Kontext des KI-Red-Teamings die **Hauptaufgabe** von KI-Sicherheitstests?

A) Die Beurteilung, ob ein KI-Modell widerstandsfähig gegenüber bestimmten Angriffen ist, indem diese Angriffe in einer kontrollierten Umgebung reproduziert werden
B) Die Bewertung, ob ein KI-Modell genaue und rückverfolgbare Ausgaben liefert, wenn es mit den erwarteten Eingabedaten konfrontiert wird
C) Die Messung, ob ein KI-Modell bei der Verarbeitung eines großen Anfrageaufkommens die erwartete Recheneffizienz aufweist
D) Die Verifizierung mittels Standardverfahren der Qualitätssicherung, dass ein KI-Modell die festgelegten geschäftlichen Anforderungen erfüllt

A) Richtig. KI-Sicherheitstests sind der Teil des KI-Red-Teamings, bei dem durch die Simulation bestimmter Angriffe geprüft wird, ob das KI-Modell diesen standhalten kann. Ihr Hauptzweck besteht daher darin, die Widerstandsfähigkeit des Modells durch die kontrollierte Reproduktion von Angriffen zu überprüfen. (Literatur: A, Kapitel 5)
B) Falsch. Die Bewertung der Ausgabegenauigkeit unter normalen Eingabebedingungen beschreibt Funktions- oder Leistungstests, nicht KI-Sicherheitstests, deren Schwerpunkt auf der Angriffssimulation liegt.
C) Falsch. Die Recheneffizienz bei routinemäßigen Arbeitslasten ist Gegenstand von Leistungs- oder Lasttests, nicht jedoch des angriffszentrierten Anwendungsbereichs von KI-Sicherheitstests.
D) Falsch. Die Überprüfung der Compliance mit geschäftlichen Anforderungen durch Qualitätssicherung liegt außerhalb des Umfangs von KI-Sicherheitstests, die speziell auf die Fähigkeit des Modells abzielen, Angriffen standzuhalten.

34 / 40
Ein Unternehmen stellt einen generativen KI-Assistenten bereit und möchte KI-Sicherheitstests definieren, die über herkömmliche Penetrationstests hinausgehen. Im ersten Schritt identifiziert das Testteam die Bedrohungen, auf die getestet werden sollte.

Welche **spezifisch** mit generativen KI-Systemen verbundenen Bedrohungen sollten in die KI-Sicherheitstests einbezogen werden?

A) Evasion Attacks (Umgehungsangriffe), unsichere Ausgabeverarbeitung und Model Poisoning (Modellvergiftung)
B) Model Exfiltration (Modellausleitung), Prompt-Injection und Evasion Attacks
C) Prompt-Injection, Offenlegung sensibler Daten und unsichere Ausgabeverarbeitung
D) Ausgabe sensibler Daten durch das Modell, Modelldrift und unsicherer Tool-Aufruf

A) Falsch. Evasion Attacks werden häufiger mit prädiktiven KI-Systemen in Verbindung gebracht, während unsichere Ausgabeverarbeitung generative KI-Systeme betreffen kann. Diese Kombination bildet die zentralen, für generative KI spezifischen Bedrohungen, die in die KI-Sicherheitstests einbezogen werden sollten, nicht am besten ab.
B) Falsch. Obwohl es sich hierbei um gültige KI-Sicherheitsbedrohungen handelt, werden Evasion Attacks häufiger mit prädiktiven KI-Systemen in Verbindung gebracht als mit generativen KI-Assistenten.
C) Richtig. Diese drei Bedrohungen betreffen generative KI und sollten daher getestet werden. (Literatur: A, Kapitel 5)
D) Falsch. Modelldrift ist ein Leistungsproblem, keine Sicherheitsbedrohung.

35 / 40
Ein KI-Sicherheitsteam testet einen generativen KI-Chatbot auf seine Widerstandsfähigkeit gegen Prompt-Injection. Während des Tests hat das System mehrere konstruierte Angriffseingaben blockiert.

Welcher Schritt wird als **Nächstes** empfohlen, um die Sicherheitsaufsicht des generativen KI-Chatbots aufrechtzuerhalten?

A) Die Eingaben beispielsweise durch den Austausch von Synonymen, Änderung der Kodierung oder Anpassung der Formatierung variieren, um zu versuchen, die Erkennung durch das KI-System zu umgehen
B) Den aktuellen Testzyklus abschließen und die Abwehrmaßnahmen als ausreichend erachten, da das System die meisten Prompt-Injection-Versuche blockiert hat
C) Denselben Test mehrmals wiederholen, um die Stichprobengröße zu erhöhen und statistisch aussagekräftigere Daten bezüglich der Abwehrleistung zu erhalten
D) Die Eingaben für Prompt-Injection-Angriffe direkt an das Modell senden, um es hinsichtlich ähnlicher Bedrohungen in zukünftigen Bereitstellungen besser zu trainieren

A) Richtig. Das Hinzufügen von Variationsalgorithmen wie Synonymen, Kodierungen und Formatierungen, um die Erkennung zu umgehen, erhöht die Robustheit der Erkennung. Die Erweiterung der Eingabevielfalt verbessert die Abdeckung, deckt Umgehungsmöglichkeiten auf und erhält die Sicherheitsaufsicht aufrecht. (Literatur: A, Kapitel 5)
B) Falsch. Der Test sollte noch nicht abgeschlossen werden. Das System muss weiterhin widerstandsfähig gegen gezielte Umgehungsversuche sein.
C) Falsch. Die erneute Durchführung derselben Tests ohne Variation wird wahrscheinlich keine neuen Umgehungstechniken aufdecken oder die Abdeckung verbessern.
D) Falsch. Das direkte Senden von rohen Angriffseingaben an das Modell umgeht Produktionsfilter und Erkennungsmechanismen. Darüber hinaus stellt das direkte Einspeisen von Angriffseingaben in das Training eine Maßnahme zur Modellverbesserung dar und ist nicht der nächste Schritt im Testverfahren.

36 / 40
Das KI-System eines Unternehmens stützt sich beim Training und bei der Entscheidungsfindung auf große Mengen personenbezogener Daten.

Um welches Risiko sollte sich das Unternehmen aus ethischer Sicht **vorrangig** kümmern?

A) Die Erhebung, Speicherung und Weiterverwendung von Daten, die über den gerechtfertigten Zweck hinausgeht
B) Die Komplexität der Benutzeroberfläche, die die Überprüfung automatisierter Entscheidungen erschwert
C) Die sinkende Genauigkeit des Modells bei ungleichmäßiger Ressourcenzuweisung
D) Softwareupdates, die plattformübergreifende Kompatibilitätsprobleme verursachen

A) Richtig. KI-Systeme sind datenintensiv und werfen Bedenken hinsichtlich der Erhebung, Aufbewahrung und Wiederverwendung personenbezogener Daten auf. Datenschutzgrundsätze und Rechtsvorschriften verlangen für eine solche Nutzung häufig eine Rechtsgrundlage oder Einwilligung sowie Schutzmaßnahmen wie Verwendungsbegrenzung und Datenminimierung. (Literatur: A, Kapitel 6)
B) Falsch. Die Komplexität der Benutzeroberfläche ist in erster Linie ein Problem der Benutzerfreundlichkeit. Sie betrifft nicht unmittelbar die Risiken, die sich aus KI-Systemen ergeben, die auf große Mengen personenbezogener Daten angewiesen sind.
C) Falsch. Die Zuweisung von Hardwareressourcen betrifft die technische Leistung und den Modellbetrieb. Sie ist nicht das Hauptanliegen im Zusammenhang mit der Verwendung personenbezogener Daten in KI-Systemen.
D) Falsch. Probleme mit Softwareupdates und Kompatibilität betreffen die Systemwartung und Interoperabilität. Sie stellen nicht das Hauptanliegen dar, das sich aus KI-Systemen ergibt, die große Mengen personenbezogener Daten verwenden.

37 / 40
Eine Online-Buchhandlung analysiert mit Hilfe eines KI-Systems Browserverlauf, Kauftrends und Vorbestellungen, um die Nachfrage vorherzusagen und sicherzustellen, dass beliebte Bücher rechtzeitig vorrätig sind. Später verwendet das Unternehmen dieselben Kundendaten erneut, um ein separates KI-Modell zu trainieren, das detaillierte Marketingprofile erstellt und gezielte Werbung an einzelne Käufer sendet.

Gegen welchen Datenschutzgrundsatz verstößt dies **am unmittelbarsten**?

A) Datenminimierung
B) Begrenzung der Datenaufbewahrung
C) Transparenz und Einwilligung
D) Verwendungsbegrenzung und Zweckbestimmung

A) Falsch. Die Datenminimierung bezieht sich darauf, nur die Daten zu erheben, die für einen definierten Zweck unbedingt erforderlich sind, und nicht auf die Einschränkung der Weiterverwendung bereits erhobener Daten für einen anderen Zweck.
B) Falsch. Die Begrenzung der Datenaufbewahrung betrifft die Dauer der Speicherung von Daten, bevor diese gelöscht werden müssen, und nicht die Frage, ob für einen bestimmten Zweck erhobene Daten für einen anderen Zweck wiederverwendet werden dürfen.
C) Falsch. Transparenz und Einwilligung beziehen sich auf die Unterrichtung betroffener Personen über die Datenerhebung und die Einholung ihrer Einwilligung, nicht speziell auf die Beschränkung, für einen bestimmten Zweck erhobene Daten zum Trainieren eines Modells für einen anderen Zweck zu verwenden.
D) Richtig. Das Szenario veranschaulicht einen Verstoß gegen die Verwendungsbegrenzung und die Zweckbestimmung, da Daten, die aus Sicherheitsgründen erhoben wurden, nicht als Trainingsdatensatz für Profiling oder personalisiertes Marketing umgenutzt werden dürfen. (Literatur: A, Kapitel 6)

38 / 40
Eine Organisation möchte ein Framework erstellen, um den verantwortungsvollen KI-Einsatz in allen ihren Geschäftsbereichen zu regeln.

Sie nutzt die Norm ISO/IEC 27001 für das Informationssicherheitsmanagement. Das Sicherheitsteam möchte seine KI-Governance anhand einer ISO/IEC-Norm aufbauen, die für die KI-Governance denselben strukturellen Zweck erfüllt wie die Norm ISO/IEC 27001 für die Informationssicherheit.

Um welche Norm handelt es sich hierbei?

A) ISO/IEC 23894
B) ISO/IEC 27005
C) ISO/IEC 42001
D) ISO/IEC 5338

A) Falsch. ISO/IEC 23894 konzentriert sich auf Leitlinien zum KI-Risikomanagement und nicht auf die Bereitstellung eines Managementsystems für die Governance verantwortungsvoller KI.
B) Falsch. ISO/IEC 27005 befasst sich mit dem Risikomanagement für Informationssicherheit und ist nicht als Governance-Managementsystem für verantwortungsvolle KI konzipiert.
C) Richtig. ISO/IEC 42001 kann als Managementsystemnorm für die Governance verantwortungsvoller KI in einer Organisation angesehen werden, ähnlich wie ISO/IEC 27001 eine Managementsystemnorm für die Informationssicherheit ist. (Literatur: A, Kapitel 1)
D) Falsch. ISO/IEC 5338 konzentriert sich auf die Erweiterung von Praktiken und Engineering im Softwarelebenszyklus und nicht auf die Bereitstellung eines Governance-Managementsystems für verantwortungsvolle KI.

39 / 40
Ein Unternehmen möchte ein KI-System einsetzen, um Personen im Bewerbungsverfahren anhand ihrer Lebensläufe automatisch zu prüfen und in eine Rangfolge zu bringen.

In welche Kategorie ist die Nutzung dieses KI-Systems laut KI-Verordnung einzuordnen?

A) Unannehmbares Risiko
B) Hohes Risiko
C) Begrenztes Risiko
D) Minimales oder kein Risiko

A) Falsch. Die KI-Verordnung verbietet KI-basierte Einstellungstools nicht. Sie ordnet diese der Kategorie hohes Risiko zu, was bedeutet, dass sie bestimmten Anforderungen und Einschränkungen unterliegen, jedoch nicht gänzlich verboten sind.
B) Richtig. Die Überprüfung von Bewerbern kann den Zugang zur Arbeit und Grundrechte beeinträchtigen und muss der Kategorie hohes Risiko zugeordnet werden. Organisationen müssen die Compliance sicherstellen, bevor sie KI-Systeme für solche Zwecke bereitstellen. (Literatur: A, Kapitel 6)
C) Falsch. Das begrenzte Risiko umfasst in erster Linie Verpflichtungen, die sich ausschließlich auf Transparenz beziehen. Die Überprüfung von Bewerbern kann den Zugang zur Arbeit und Grundrechte beeinträchtigen und muss der Kategorie hohes Risiko zugeordnet werden.
D) Falsch. Ein minimales Risiko gilt für Anwendungen ohne weitreichende Folgen, wie beispielsweise Spamfilter oder Videospiele. Die Überprüfung von Bewerbern kann den Zugang zur Arbeit und Grundrechte beeinträchtigen und muss der Kategorie hohes Risiko zugeordnet werden.

40 / 40
Ein KI-Entwicklungsteam bereitet die Trainingsdaten für ein neues Modell für maschinelles Lernen vor. Der Teamleiter möchte sicherstellen, dass das Vorgehen bei der Datenbeschaffung die Compliance mit dem Urheberrecht (Copyright Compliance) unterstützt.

Welche Vorgehensweise bei der Datenbeschaffung ist **am ethischsten**?

A) Daten ausschließlich aus öffentlich zugänglichen Quellen zu erheben
B) Trainingsdaten ausschließlich organisationsintern zu erstellen
C) Daten mit stillschweigender Genehmigung aus Quellen herunterzuladen
D) Ausschließlich Daten zu verwenden, die wiederholt online weiterverbreitet wurden

A) Falsch. Die bloße öffentliche Verfügbarkeit garantiert nicht, dass die Daten für die von der Organisation beabsichtigte Nutzung lizenziert sind. Die Lizenzbedingungen müssen überprüft werden, um sicherzustellen, dass sie für die geplante Anwendung ausreichend sind.
B) Richtig. Die interne Erstellung von Daten ist eine der anerkannten ethischen Strategien zur Datenbeschaffung, da sie sicherstellt, dass die Organisation die vollen Rechte an den Daten besitzt, die zum Training von KI-Modellen verwendet werden. (Literatur: A, Kapitel 0)
C) Falsch. Eine stillschweigende Genehmigung ist keine gültige Grundlage für die Datennutzung. Vor der Verwendung von Daten Dritter für das KI-Training müssen ausdrückliche Genehmigungen vom Dateneigentümer eingeholt werden.
D) Falsch. Inhalte, die online vielfach erneut veröffentlicht werden, können dennoch durch das Urheberrecht geschützt sein. Die Wiederholung oder weite Verbreitung macht das Material nicht frei für die Wiederverwendung zum KI-Training. Der Fokus muss darauf liegen, ob die Daten intern erstellt, ordnungsgemäß lizenziert oder mit den erforderlichen Genehmigungen beschafft wurden.

Beurteilung

Die richtigen Antworten auf die Fragen in dieser Musterprüfung finden Sie in nachstehender Tabelle.

Frage	Antwort	Frage	Antwort
1	A	**21**	B
2	A	**22**	A
3	A	**23**	C
4	D	**24**	B
5	A	**25**	C
6	C	**26**	A
7	E	**27**	C
8	A	**28**	B
9	C	**29**	D
10	A	**30**	D
11	C	**31**	C
12	C	**32**	C
13	B	**33**	A
14	A	**34**	C
15	B	**35**	A
16	A	**36**	A
17	A	**37**	D
18	D	**38**	C
19	C	**39**	B
20	C	**40**	B

Certified for what's next

Kontakt EXIN

www.exin.com