

Exemple d'examen

Édition 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.

Table des matières

Introduction	4
Exemple d'examen	5
Solutions à l’examen	22
Évaluation	60

Introduction

Voici l'exemple d'examen EXIN AI Security Professional based on the OWASP AI Exchange (AISP.FR). Les règles et réglementations d'examens EXIN s'appliquent à cet examen.

Cet examen consiste en 40 questions à choix multiples. Chaque question à choix multiple comporte un certain nombre de réponses possibles dont seulement une est correcte.

Le maximum de points qui peut être obtenu lors de l'examen est de 40. Chaque réponse correcte rapporte 1 point. Si vous obtenez 26 points ou plus vous réussissez votre examen.

Le temps alloué lors de l'examen est de 90 minutes.

Bonne chance !

Exemple d'examen

1 / 40
Un responsable de la sécurité informatique travaille dans une entreprise de logistique. L'entreprise a récemment décidé d'adopter des outils d'IA pour optimiser ses itinéraires de livraison. Le responsable de la sécurité informatique dirige les efforts visant à mettre en place la sécurité de l'IA. Les deux premières étapes du cadre G.U.A.R.D. sont déjà achevées :

- L'étape « Govern » (Gouverner) a été réalisée en mettant en place une politique de sécurité de l'IA.
- L'étape « Understand » (Comprendre) a été réalisée en répertoriant tous les actifs liés à l'IA et leurs risques.

Quelle est la **prochaine** étape que le responsable de la sécurité informatique doit entreprendre ?

A) Adapter les pratiques de sécurité pour y inclure la modélisation des menaces spécifiques à l'IA, les tests et les contrôles de la chaîne d'approvisionnement
B) Analyser les risques liés à l'IA identifiés et établir une stratégie de hiérarchisation des risques en fonction de leur impact potentiel
C) Appliquer des contrôles de sécurité à l'ensemble des systèmes d'IA afin d'instaurer une sécurité de l'IA responsable tout au long de son cycle de vie
D) Évaluer la politique de sécurité de l'IA élaborée lors de l'étape « Govern » afin d'y inclure les actifs liés à l'IA répertoriés et leurs risques

2 / 40
Une entreprise examine sa stratégie en matière d'IA afin de s'assurer qu'elle profite à l'activité tout en étant éthique et solide sur le plan opérationnel. Le responsable de la sécurité de l'IA classe les deux dimensions principales de l'IA responsable et de l'IA digne de confiance.

Quelle est la **principale** différence entre l'IA responsable et l'IA digne de confiance ?

A) - L'IA responsable met l'accent sur l'éthique, la société et la gouvernance.
- L'IA digne de confiance met l'accent sur les aspects techniques et opérationnels, la transparence et l'explicabilité.
B) - L'IA responsable met l'accent sur la conformité réglementaire aux lois et législations applicables.
- L'IA digne de confiance met l'accent sur la protection des données et la cybersécurité afin de respecter les principes de protection de la vie privée.
C) - L'IA responsable met l'accent sur l'exactitude du système et les indicateurs de performance technique.
- L'IA digne de confiance met l'accent sur les lignes directrices éthiques et l'impact social afin de susciter la confiance des consommateurs.
D) - L'IA responsable met l'accent sur la sélection des fournisseurs, les risques et la gestion des coûts.
- L'IA digne de confiance met l'accent sur une formation des utilisateurs sûre et fiable ainsi que sur l'adoption des systèmes d'IA.

3 / 40
Une équipe de cybersécurité dispose déjà d'un cadre de détection des menaces. Elle doit désormais également protéger un nouveau grand modèle de langage (LLM) utilisé dans le service client pour un agent conversationnel (chatbot).

L'équipe souhaite utiliser les mêmes outils que ceux qu'elle utilise pour les applications web :
- Règles de détection basées sur les signatures
- Contrôles fixes de validation des entrées

Pourquoi cette approche **n'est-elle pas** suffisante pour sécuriser le LLM ?

A) Parce que les systèmes d'IA présentent des surfaces d'attaque différentes et nécessitent des contrôles spécifiques pour le comportement du LLM et son pipeline de données
B) Parce que le cadre de détection des menaces n'inclut pas de procédures de réponse aux incidents, ce qui le rend inadapté aux LLM
C) Parce que les LLM génèrent des sorties trop variables pour être évaluées par des systèmes basés sur des règles et doivent être examinés par un humain
D) Parce que les LLM traitent les requêtes plus rapidement que les applications web, ce qui signifie que les règles basées sur les signatures ne peuvent pas suivre leur rythme
E) Parce que les pare-feu d'applications web ne sont pas sous licence pour une utilisation avec les LLM et nécessitent des contrats distincts avec les fournisseurs

4 / 40
Lorsqu'elle utilise l'IA, une entreprise dispose d'actifs spécifiques à l'IA. Chaque actif spécifique à l'IA présente également ses propres menaces clés.

Quelle combinaison **correcte** associe un actif spécifique à l'IA à sa menace clé ?

A) - Actif : journaux d'audit générés par les pare-feu réseau
- Menace : épuisement des ressources d'IA
B) - Actif : identifiants d'authentification des utilisateurs finaux
- Menace : sortie contenant une injection conventionnelle
C) - Actif : code source logiciel stocké dans un système de contrôle de version
- Menace : empoisonnement direct du modèle en phase d'exécution
D) - Actif : données d'entraînement utilisées pour construire un modèle
- Menace : divulgation de données sensibles dans la sortie

5 / 40
Un cabinet d'avocats utilise un système d'IA pour l'examen et l'analyse de documents. L'équipe de sécurité applique une approche de gestion des risques. Elle a dressé un inventaire de toutes les menaces potentielles pesant sur le système d'IA.

Quelle est la **prochaine** étape que l'équipe devrait entreprendre ?

A) Appliquer un processus de modélisation des menaces afin de traduire ces menaces en un ensemble de risques concrets et hiérarchisés.
B) Désigner un responsable des risques pour chaque menace identifiée afin de garantir une protection par conception depuis le début.
C) Déployer des contrôles de sécurité pour chaque menace afin d'atténuer les risques supplémentaires tout au long du cycle de vie du système d'IA.
D) Partager la liste complète des menaces avec les parties prenantes concernées afin de les sensibiliser et d'assurer la transparence.

6 / 40
Un ingénieur en sécurité de l'IA examine la conception du contrôle d'accès d'un système d'IA multi-agents. Chaque agent du système est doté d'un compte de service partagé. Ce compte dispose d'un accès en lecture et en écriture à la base de données de production, au système de fichiers interne et à la passerelle API (interface de programmation d'application) externe.

Les différents agents IA invoquent de manière autonome des outils et exécutent des workflows en plusieurs étapes sur l'ensemble des systèmes internes. La mise à disposition a été effectuée sur un compte partagé afin de simplifier la gestion des identifiants tout au long du pipeline des agents IA.

Pourquoi cette conception de compte partagé est-elle **particulièrement risquée** dans un contexte d'IA agentique ?

A) Parce que les agents IA génèrent des volumes d'appels API plus importants que les utilisateurs humains, ce qui augmente le risque que les contrôles de limitation de débit soient déclenchés et que le compte partagé soit suspendu
B) Parce que les agents IA n'utilisent pas d'authentification multifactorielle ; les identifiants partagés constituent donc un point de défaillance unique, sans autre couche de vérification, qu'un attaquant peut exploiter
C) Parce que les agents IA opèrent de manière autonome à travers plusieurs étapes de raisonnement au cours d'une même session ; un agent compromis peut donc enchaîner des actions sur tous les systèmes accessibles
D) Parce que les cadres réglementaires applicables aux systèmes d'IA exigent que chaque agent IA soit doté d'identifiants uniques ; les comptes partagés constituent donc une violation directe de la conformité
E) Parce que le partage de comptes de service entre plusieurs agents IA engendre une complexité opérationnelle excessive lors de la rotation obligatoire des identifiants et des activités de gestion des accès

7 / 40
Un ingénieur en apprentissage automatique au sein d'une institution financière enquête sur deux incidents distincts impliquant le modèle de détection des fraudes de l'institution.

- **Incident 1** : un attaquant externe, n'ayant accès ni à l'architecture du modèle, ni aux données d'entraînement, ni aux paramètres, soumet des milliers de transactions spécialement conçues. Celles-ci sont affinées de manière itérative en fonction du déclenchement ou non de l'alerte de fraude.
- **Incident 2** : un chercheur utilise un modèle de substitution entraîné localement, doté d'une architecture similaire, pour générer des transactions adversariales. Celles-ci parviennent à tromper le modèle de production sans interaction préalable.

De quelles deux stratégies d'évasion s'agit-il ?

A) - Incident 1 : évasion après empoisonnement
- Incident 2 : évasion à connaissance parfaite

B) - Incident 1 : évasion à connaissance partielle
- Incident 2 : attaque par transfert

C) - Incident 1 : évasion à connaissance parfaite
- Incident 2 : évasion à connaissance nulle

D) - Incident 1 : attaque par transfert
- Incident 2 : évasion à connaissance partielle

E) - Incident 1 : évasion à connaissance nulle
- Incident 2 : attaque par transfert

8 / 40
Erick est un utilisateur d'une plateforme de réseaux sociaux qui modère automatiquement les commentaires à l'aide d'un système d'IA. Il tente de faire passer un commentaire enfreignant les règles à travers la modération automatisée de la plateforme, car il souhaite stimuler l'engagement sur sa publication. Erick ne dispose d'aucune information sur l'architecture du modèle, les paramètres ou les données d'entraînement, et il ne voit qu'un message binaire lorsqu'il tente de publier son commentaire : soit « bloqué », soit « publié ». En apportant à plusieurs reprises de légères modifications à la formulation de son commentaire, Erick finit par trouver une version qui est publiée.

Quel type de menace d'évasion liée aux données d'entrée est décrit ici ?

A) Évasion à connaissance nulle
B) Évasion à connaissance partielle
C) Évasion à connaissance parfaite
D) Évasion après empoisonnement
E) Attaque par transfert

9 / 40
Une société de services juridiques développe un agent conversationnel (chatbot) capable de répondre aux questions des utilisateurs et d'extraire des informations à partir de ressources internes et externes, telles que des sites web et des documents d'entreprise. Lors d'une démonstration, le comportement de l'agent conversationnel déclenche une enquête menée par l'ingénieure en sécurité de l'IA. Elle identifie deux menaces distinctes :

- **Menace 1 :** un message conçu pour passer outre les instructions et les règles de sécurité de l'agent conversationnel peut amener ce dernier à suivre les commandes du testeur plutôt que la politique que sa configuration lui impose de suivre.
- **Menace 2 :** une page web d'un fournisseur récupérée par l'agent conversationnel contient du texte intégré qui agit comme des instructions. Lorsque l'agent conversationnel ingère cette page, il traite le contenu comme des commandes et s'écarte de sa politique.

De quels types de menaces s'agit-il ?

A) - Menace 1 : évasion
- Menace 2 : empoisonnement des données

B) - Menace 1 : empoisonnement des données
- Menace 2 : injection de prompt directe

C) - Menace 1 : injection de prompt directe
- Menace 2 : injection de prompt indirecte

D) - Menace 1 : injection de prompt indirecte
- Menace 2 : évasion

10 / 40
Une société de conseil dispose d'un système d'IA destiné à faciliter ses tâches quotidiennes. À l'issue d'une réunion avec un client, l'assistant IA rédige un résumé de l'entretien. L'assistant IA commence à insérer des liens internes confidentiels dans le brouillon d'e-mail destiné au client. Cependant, avant l'envoi, un détecteur d'anomalies autonome suspend l'envoi et met le brouillon en quarantaine.

L'architecte de la sécurité de l'IA souhaite limiter à l'avance l'accès de l'assistant IA chargé des réunions aux seuls droits des personnes concernées.

Quelles couches de protection contre l'injection de prompt sont illustrées ici ?

A) Contrôle automatisé et moindre privilège fondé sur l'utilisateur
B) Autorisation « juste à temps » et contrôle humain
C) Alignement du modèle et moindre privilège fondé sur l'intention
D) Gestion des entrées/sorties face à l'injection de prompt et limitation du débit

11 / 40
Un auditeur évalue les risques liés à la protection de la vie privée d'un grand modèle de langage (LLM) déployé dans le secteur de la santé, ainsi que sa conformité aux lois sur la protection des données. Pour ce faire, l'auditeur interroge le modèle à l'aide de requêtes ciblées concernant un patient spécifique et des non-patients synthétiques soigneusement appariés. En observant les différences dans les schémas de réponse et les scores de confiance du modèle, l'auditeur peut déterminer de manière fiable si le dossier de ce patient a été inclus dans les données d'entraînement, même si aucun détail médical réel n'est révélé.

Quel type d'attaque est simulé ici ?

A) Divulgation de données dans la sortie du modèle
B) Injection de prompt indirecte
C) Inférence d'appartenance
D) Inversion de modèle

12 / 40
Quelle est l'une des **principales** conséquences d'une attaque d'exfiltration du modèle réussie ?

A) L'attaquant peut extraire l'ensemble de données d'entraînement brut utilisé pour construire le modèle d'origine, qui peut potentiellement contenir des enregistrements sensibles.
B) L'attaquant prend le contrôle direct du modèle déployé et peut modifier ses poids et ses paramètres sans autorisation préalable.
C) L'attaquant obtient une réplique du modèle qui peut être utilisée pour orchestrer des attaques d'évasion sans déclencher les défenses de l'original.
D) Le modèle d'origine devient indisponible pour les utilisateurs légitimes en raison de l'épuisement des ressources, ce qui entraîne une dégradation du service et des délais d'attente répétés.

13 / 40
Un grand modèle de langage (LLM) hébergé dans le cloud est victime d'une attaque éponge. Les attaquants envoient des entrées très longues ou complexes qui imposent un surcroît de calcul, ce qui fait grimper les coûts et provoque des ralentissements ou des pannes. Afin de remédier à cette situation, l'ingénieur en sécurité de l'IA souhaite mettre en place deux contrôles :

- **Le contrôle 1** validera et assainira les entrées afin de rejeter ou de corriger les contenus malveillants, tels que les entrées anormalement volumineuses.
- **Le contrôle 2** limitera les ressources de calcul consommées par entrée du modèle afin d'empêcher toute surutilisation.

Quels sont ces deux contrôles ?

A) - Contrôle 1 : gestion des entrées anormales
- Contrôle 2 : masquage des limites de confiance
B) - Contrôle 1 : validation des entrées contre le déni de service (DoS)
- Contrôle 2 : limitation des ressources
C) - Contrôle 1 : contrôle d'accès au modèle
- Contrôle 2 : modèle résistant à l'évasion
D) - Contrôle 1 : surveillance de l'utilisation
- Contrôle 2 : limitation de débit

14 / 40
Un ingénieur en sécurité de l'IA identifie les menaces suivantes pour une entreprise :

- **Des prompts adversariaux** sont conçus par des attaquants pour provoquer des erreurs de classification.
- Un **pipeline d'ingestion compromis** introduit des enregistrements contenant des portes dérobées dans le corpus avant la construction du modèle d'IA.
- Les attaquants altèrent les **workflows d'étiquetage** en modifiant les annotations de référence afin de créer un biais dans le comportement du modèle d'IA.
- Les **données d'entraînement** proviennent de sources externes ou tierces sans validation appropriée.

Selon l'OWASP AI Exchange, quelle menace **n'implique pas** de risques d'empoisonnement des données (phase de développement) ?

A) Entrées adversariales
B) Pipeline d'ingestion compromis
C) Workflows d'étiquetage
D) Données d'entraînement

15 / 40
Selon l'OWASP AI Exchange, quel exemple illustre un **empoisonnement direct du modèle en phase de développement** ?

A) Les attaquants créent des entrées adversariales en phase d'exécution pour induire le modèle en erreur
B) Les attaquants manipulent les poids et le pipeline d'entraînement pour modifier le comportement du modèle.
C) Des données d'entraînement sensibles sont involontairement exposées via les sorties du modèle.
D) Un pic de trafic submerge le service d'IA, entraînant un ralentissement ou une interruption de service.

16 / 40
Selon l'OWASP AI Exchange, quel exemple illustre un **empoisonnement de modèle dans la chaîne d'approvisionnement** ?

A) Les attaquants compromettent les données, les modèles ou les composants de la chaîne d'approvisionnement avant leur intégration.
B) Les attaquants manipulent les données d'entrée lors de la phase d'exécution afin d'induire le modèle en erreur et de lui faire produire des sorties erronées.
C) Le modèle produit des résultats erronés en raison d'une qualité insuffisante des données ou de l'entraînement.
D) Le modèle divulgue involontairement des données d'entraînement sensibles par le biais de ses réponses.

17 / 40
Une équipe de développement en IA procède à une évaluation des risques de son environnement de développement. Elle identifie deux menaces distinctes :

La menace 1 implique un accès non autorisé aux ensembles de données utilisés pour entraîner et tester son modèle d'IA.
La menace 2 implique un accès non autorisé aux paramètres et aux poids du modèle stockés dans un référentiel compromis.

De quels types de menaces s'agit-il ?

A) - Menace 1 : fuite de données en phase de développement
- Menace 2 : fuite directe du modèle en phase de développement

B) - Menace 1 : fuite directe des données d'augmentation
- Menace 2 : fuite du référentiel

C) - Menace 1 : fuite directe de données en phase d'exécution
- Menace 2 : fuite directe du modèle en phase d'exécution

D) - Menace 1 : fuite des données d'entrée
- Menace 2 : fuite du code source ou de la configuration

18 / 40
Une équipe de sécurité examine le paysage des menaces de sécurité en phase d'exécution d'un système d'IA. Ce système d'IA est intégré à une application web. L'équipe identifie deux menaces de sécurité conventionnelles :

- **Menace 1 :** un attaquant peut accéder à la mémoire du serveur de production et extraire les poids et les paramètres du modèle.
- **Menace 2 :** un attaquant ayant accès au registre des modèles remplace l'artefact du modèle déployé par une version contenant un cheval de Troie.

Quelles sont les conséquences de ces menaces de sécurité conventionnelles qui sont **spécifiques** aux systèmes d'IA ?

A) - La menace 1 entraîne le réentraînement automatique du système sur des données malveillantes.
- La menace 2 force le modèle à rejeter des prompts valides.

B) - La menace 1 modifie l'architecture du modèle en phase d'exécution.
- La menace 2 compresse silencieusement les poids pour réduire leur précision.

C) - La menace 1 désactive les filtres de sécurité, rendant le système plus permissif de par sa conception.
- La menace 2 rend les sorties déterministes et prévisibles.

D) - La menace 1 permet le vol de modèles et l'inférence potentielle des données d'entraînement.
- La menace 2 permet l'altération non détectée des modèles.

19 / 40
Une entreprise de services financiers déploie un modèle d'apprentissage automatique (ML) pour détecter les transactions frauduleuses. Le modèle ML est hébergé dans son environnement de production et traite en continu des données de transactions en temps réel. Deux incidents de sécurité se produisent :

- **Incident 1** : un attaquant obtient un accès non autorisé à l'environnement du modèle déployé et modifie les paramètres du modèle de sorte que certaines transactions frauduleuses ne soient plus signalées. Cette manipulation n'est pas détectée, et l'attaquant en tire profit.
- **Incident 2** : un attaquant exploite une vulnérabilité du système de production. Cela lui permet de répliquer le modèle et d'analyser son comportement en vue d'autres attaques.

D'après l'OWASP AI Exchange, de quel type d'incidents de sécurité s'agit-il ?

A) - Incident A : empoisonnement des données
- Incident B : empoisonnement direct du modèle en phase d'exécution

B) - Incident A : empoisonnement des données
- Incident B : attaque d'évasion

C) - Incident A : empoisonnement direct du modèle en phase d'exécution
- Incident B : fuite directe du modèle en phase d'exécution

D) - Incident A : fuite directe du modèle en phase d'exécution
- Incident B : attaque d'évasion

20 / 40
Un établissement de santé déploie un assistant IA pour aider les cliniciens à synthétiser les dossiers des patients et à formuler des recommandations. Le système s'intègre à des bases de données internes et à des outils externes, et les cliniciens s'appuient sur les sorties qu'il génère dans le cadre de leurs activités quotidiennes. Deux menaces sont observées :

- **Menace 1** : l'assistant IA génère une réponse contenant un script malveillant caché intégré dans du contenu HTML. Lorsque la sortie est consultée, le script s'exécute et envoie les données de session à l'attaquant.
- **Menace 2** : un attaquant exploite des contrôles d'accès insuffisants dans le système pour intercepter et récupérer des données sensibles sur les patients.

D'après l'OWASP AI Exchange, quels types de menaces sont décrits ?

A) - Menace 1 : empoisonnement des données
- Menace 2 : fuite des données d'entrée

B) - Menace 1 : empoisonnement des données
- Menace 2 : attaque par inversion de modèle

C) - Menace 1 : sortie contenant une injection conventionnelle
- Menace 2 : fuite des données d'entrée

D) - Menace 1 : sortie contenant une injection conventionnelle
- Menace 2 : attaque par inversion de modèle

21 / 40
Une entreprise utilise un système de génération augmentée par récupération (RAG) pour répondre aux questions de ses employés concernant les politiques internes. Le système d'IA extrait des documents d'une base de données vectorielle et les ajoute au prompt avant de les envoyer au modèle d'IA.

Un audit de sécurité révèle qu'un attaquant a obtenu un accès non autorisé à la base de données vectorielle. L'attaquant a consulté les documents d'augmentation stockés et a remplacé un document de politique légitime par un autre, contenant de fausses informations. Les employés qui interrogent le système d'IA au sujet de cette politique reçoivent désormais des conseils erronés.

Quelles menaces de sécurité ce scénario illustre-t-il ?

A) - Menace 1 : fuite des données d'augmentation
- Menace 2 : sortie contenant une injection conventionnelle

B) - Menace 1 : manipulation des données d'augmentation
- Menace 2 : fuite des données d'augmentation

C) - Menace 1 : inférence d'appartenance
- Menace 2 : manipulation des données d'augmentation

D) - Menace 1 : sortie contenant une injection conventionnelle
- Menace 2 : inférence d'appartenance

22 / 40
Des contrôles de gouvernance généraux doivent être mis en œuvre pour garantir une supervision efficace de la sécurité des systèmes d'IA.

Quelle approche l'illustre le **mieux** ?

A) Définir des politiques, des rôles et des pratiques de gestion des risques clairs pour guider le développement, le déploiement et la surveillance sécurisés des systèmes d'IA

B) Intégrer des contrôles de sécurité dans les systèmes d'IA afin d'appliquer les principes de « sécurité dès la conception » lors de la phase de conception du système d'IA

C) Respecter les exigences minimales de sécurité, car les systèmes d'IA sont plus sûrs par défaut et ne nécessitent pas de multiples couches de sécurité

D) Séparer les contrôles de gouvernance de l'IA des contrôles de sécurité de l'information afin de s'assurer que ces deux concepts distincts soient traités de manière équivalente

23 / 40
Une agence de marketing de taille moyenne dispose de peu d'expérience en matière d'IA et n'a pas encore mis en place de gouvernance formelle de l'IA. Elle souhaite utiliser un outil d'IA générative pour rédiger des e-mails destinés à ses clients. Cet outil nécessite d'accéder aux noms et adresses e-mail des clients. L'entreprise devrait mettre en place une gouvernance de l'IA et établir une supervision de la sécurité avant de pouvoir utiliser cet outil d'IA.

Quel est le **strict minimum** que l'entreprise devrait faire en matière de supervision de la sécurité ?

A) Réaliser des exercices de red teaming sur le système d'e-mails basé sur l'IA générative afin d'atténuer les risques de sécurité avant son lancement
B) Recruter un spécialiste de la gouvernance de l'IA pour réaliser une évaluation des menaces et définir des politiques et des contrôles pour le déploiement
C) Dresser un inventaire des utilisations prévues de l'IA et réaliser une analyse des risques afin d'identifier les menaces, les contrôles et les responsabilités
D) Solliciter l'avis juridique d'un cabinet spécialisé dans l'IA et les exigences du règlement sur l'IA qui s'y rapportent
E) Commencer par un projet pilote et surveiller les performances, les utilisations abusives et les incidents du système de messagerie basé sur l'IA à l'aide de journaux

24 / 40
Quelle est la **meilleure** description de la portée des contrôles de gouvernance généraux en matière de sécurité de l'IA ?

A) Un ensemble de méthodes de chiffrement utilisées pour sécuriser les sorties de l'IA après leur génération
B) Un ensemble de contrôles globaux s'appliquant à toutes les menaces liées à l'IA et à toutes les étapes de son cycle de vie
C) Des contrôles limités à la phase d'entraînement du modèle, axés sur la qualité des ensembles de données et les techniques de réglage
D) Des mesures facultatives que les organisations peuvent adopter si les systèmes d'IA ne traitent pas de données sensibles
E) Des mesures de protection techniques appliquées exclusivement en phase d'exécution pour détecter les entrées malveillantes

25 / 40
Une entreprise du secteur de la santé décide d'utiliser un modèle d'IA prêt à l'emploi, fourni par un tiers, pour une application. L'application traite les données saisies par l'utilisateur et récupère des données externes.

L'ingénieur en sécurité de l'IA s'inquiète de la sécurité du modèle. La stratégie de sécurité définissant la répartition des responsabilités en matière de sécurité entre le fournisseur de modèle d'IA et l'entreprise du secteur de la santé n'est pas établie.

Qu'est-ce qui est **correct** concernant ces responsabilités ?

A) La plupart des contrôles de sécurité peuvent être délégués au fournisseur lors de l'utilisation de modèles d'IA prêts à l'emploi, car l'entraînement et l'hébergement assurés par le fournisseur couvrent les risques de sécurité.
B) Le déployeur est responsable de tous les contrôles de sécurité lorsqu'il intègre le modèle d'IA dans ses systèmes et tient compte de la sécurité avant le déploiement.
C) Le fournisseur a entraîné le modèle d'IA ; il gère donc les contrôles au niveau du modèle, tandis que le déployeur doit gérer les contrôles au niveau de l'application.
D) Le fournisseur est responsable de tous les contrôles de sécurité car il a entraîné le modèle d'IA et l'héberge ; par conséquent, tous les risques et toutes les attaques devraient être gérés par le fournisseur.

26 / 40
Une organisation a déployé une application d'IA destinée aux clients utilisant un grand modèle de langage (LLM) fourni par un fournisseur tiers. L'accès au modèle s'effectue via un service d'API (interface de programmation d'application) géré et celui-ci n'a pas fait l'objet d'un entraînement supplémentaire ni d'une modification de la part de l'organisation.

L'équipe de sécurité découvre que des entrées utilisateur soigneusement élaborées peuvent amener le modèle à ignorer les limites comportementales qui lui ont été configurées et à produire des sorties qui enfreignent les restrictions de contenu de l'organisation. L'équipe examine quel contrôle mettre en œuvre qui relève de la responsabilité de l'organisation en tant que déployeur plutôt que de celle du fournisseur tiers.

Quel contrôle l'équipe de sécurité devrait-elle mettre en œuvre pour résoudre ce problème ?

A) Insérer une couche de validation des sorties qui analyse toutes les réponses du modèle avant qu'elles n'atteignent les utilisateurs finaux et bloque les sorties qui enfreignent la politique de contenu de l'organisation
B) Modifier le pipeline d'inférence du fournisseur afin d'imposer des contraintes de génération plus strictes qui limitent la génération de réponses et empêchent l'apparition de contenus non autorisés
C) Demander au fournisseur de réentraîner et d'ajuster le modèle fourni en matière de sécurité à l'aide de données supplémentaires de refus et d'alignement sur la politique, afin d'améliorer la résistance face aux entrées visant à contourner les restrictions (jailbreak)
D) Exiger du fournisseur qu'il révise et applique des règles de modération de contenu renforcées à l'échelle du service, ainsi que des mesures d'application, afin que les sorties dangereuses soient filtrées globalement avant leur diffusion

27 / 40
Une équipe prépare des données d'entraînement pour un assistant d'IA générative. Elle constate qu'elle n'a pas besoin de tous les champs et de toutes les données qu'elle utilise actuellement pour garantir les performances du modèle d'IA. Dans un deuxième temps, elle envisage des contrôles visant à limiter les données sensibles et à améliorer la confidentialité et l'intégrité.

Quelle est la **meilleure** façon d'améliorer la confidentialité et l'intégrité ?

A) Conserver tous les champs dans les données d'entraînement et chiffrer les données au repos et dans les sauvegardes
B) Conserver les identifiants dans les données d'entraînement afin de simplifier les futures tâches et processus de réentraînement
C) Supprimer les champs et les données inutiles à l'entraînement et aux performances du modèle
D) N'utiliser que des données synthétiques pour toutes les phases d'entraînement et d'évaluation de l'IA générative

28 / 40
Une équipe produit entraîne un modèle d'IA destiné au service client. L'ensemble de données d'entraînement actuel comprend les adresses complètes des clients, leurs numéros de téléphone et des enregistrements hérités qui ne sont pas nécessaires pour atteindre les objectifs de performance, mais qui pourraient s'avérer utiles à l'avenir. Après qu'un exercice de red teaming a mis en évidence un risque de fuite et de manipulation des données via des détails mémorisés, l'équipe décide de réduire l'exposition. Pour ce faire, elle supprime les champs inutiles et des enregistrements entiers afin que les informations sensibles ne figurent pas dans l'ensemble de données.

Quel contrôle **général** a été mis en œuvre pour limiter les données sensibles dans cette situation ?

A) Données autorisées
B) Minimisation des données
C) Obfuscation des données d'entraînement
D) Conservation de courte durée

29 / 40
Une entreprise de vente au détail décide de ne plus stocker les informations de paiement de ses clients une fois chaque transaction terminée.

En quoi cette mesure **réduit**-elle le risque d'exposition des données ?

A) En éliminant la nécessité de chiffrer les données en transit ou archivées à des fins d'audit
B) En garantissant la conformité à toutes les réglementations applicables en matière de protection des données, ce qui réduit la complexité
C) En garantissant qu'aucun utilisateur non autorisé ne puisse accéder au réseau et à la base de données de l'entreprise
D) En empêchant que les données conservées ne fassent l'objet de fuites, ne soient reconstituées ou déduites à partir du système

30 / 40
Un réseau hospitalier régional utilise un agent IA pour lire les e-mails entrants des fournisseurs, rédiger des réponses et les envoyer automatiquement. Lors de tests de sécurité menés à la suite d'un récent incident d'hameçonnage, les testeurs ont découvert qu'une injection de prompt pouvait amener l'agent IA à ignorer ses règles et à envoyer des réponses non autorisées.

Quelle décision de conception est la **plus efficace** pour limiter les effets de ce comportement indésirable ?

A) Désactiver toutes les intégrations de l'agent IA et revenir à des workflows entièrement manuels pour la plupart des processus
B) Renforcer la formation des utilisateurs à la sécurité des prompts et mettre l'accent sur la supervision lors de l'envoi d'e-mails assisté par l'agent
C) Conserver des autorisations d'envoi autonomes étendues et passer à un modèle génératif plus avancé
D) Limiter les autorisations de l'agent IA à des périmètres spécifiques à certaines tâches et exiger une validation pour les actions à haut risque

31 / 40
Quelle est la **meilleure** façon de minimiser les effets d'un comportement indésirable du modèle ?

A) Mettre l'accent sur le renforcement des prompts, l'accès basé sur les rôles et les audits programmés des sorties du modèle
B) S'appuyer sur des filtres automatisés des sorties, la journalisation des exceptions et des contrôles ponctuels de l'équité
C) Rechercher les biais indésirables et appliquer des contrôles tels que la supervision, le moindre privilège et la validation continue
D) Recourir à une validation rigoureuse des données d'entrée, à des autorisations d'outils restreintes et à des revues périodiques impliquant une intervention humaine

32 / 40
Limiter les comportements indésirables des modèles est une stratégie essentielle pour limiter l'impact et contrôler le rayon d'impact.

Comment cette stratégie **améliore-t-elle** les performances et **réduit-elle** le risque ?

A) En se concentrant sur les attaques susceptibles d'affecter les systèmes d'IA et en limitant la supervision des facteurs de performance et de l'exactitude du système
B) En mettant en œuvre des contrôles qui éliminent la nécessité de tests et de validations continus au fil du temps afin de réduire la complexité
C) En améliorant la fiabilité opérationnelle et l'efficacité des ressources tout en minimisant les incidents et le gaspillage de requêtes de calcul
D) En augmentant l'autonomie du modèle et en lui permettant de générer davantage de sorties afin d'améliorer en continu les performances du système

33 / 40
Dans le cadre plus large du red teaming (exercice de test adversarial) de l'IA, quel est l'objectif **principal** des tests de sécurité de l'IA ?

A) Évaluer si un modèle d'IA est résilient face à des attaques spécifiques en reproduisant ces attaques dans un environnement contrôlé
B) Évaluer si un modèle d'IA produit des sorties précises et traçables lorsqu'il est exposé à des données d'entrée attendues
C) Mesurer si un modèle d'IA fait preuve de l'efficacité de calcul attendue lors du traitement de grands volumes de requêtes
D) Vérifier si un modèle d'IA répond aux exigences métier définies par le biais de procédures standard d'assurance qualité

34 / 40
Une entreprise déploie un assistant d'IA générative et souhaite définir des tests de sécurité de l'IA qui aillent au-delà des simples tests d'intrusion conventionnels. Pour ce faire, l'équipe de test commence par identifier les menaces à tester.

Quelles menaces **spécifiquement** associées aux systèmes d'IA générative devraient être incluses dans les tests de sécurité de l'IA ?

A) Attaques par évasion, gestion non sécurisée des sorties et empoisonnement du modèle
B) Exfiltration du modèle, injection de prompt et attaques par évasion
C) Injection de prompt, divulgation de données sensibles et gestion non sécurisée des sorties
D) Sortie de données sensibles par le modèle, dérive du modèle et invocation non sécurisée d'outils

35 / 40
Une équipe de sécurité de l'IA teste la résistance à l'injection de prompt d'un agent conversationnel (chatbot) d'IA générative. Au cours du test, le système a bloqué plusieurs entrées d'attaque spécialement conçues.

Quelle est la **prochaine** étape recommandée pour continuer à assurer la supervision de la sécurité de l'agent conversationnel d'IA générative ?

A) Ajouter des variations dans les entrées, telles que le remplacement par des synonymes, des modifications d'encodage ou des changements de mise en forme, afin de tenter de contourner la détection
B) Conclure le cycle de test en cours et considérer que les défenses sont adéquates, puisque la plupart des tentatives d'injection de prompt ont été bloquées
C) Répéter le même test plusieurs fois afin d'augmenter la taille de l'échantillon et d'obtenir des données plus pertinentes sur le plan statistique concernant les performances des défenses
D) Envoyer les entrées d'attaque par injection de prompt directement au modèle afin de mieux l'entraîner contre des menaces similaires lors de futurs déploiements

36 / 40
Le système d'IA d'une entreprise s'appuie sur de grandes quantités de données à caractère personnel pour l'entraînement et la prise de décision.

D'un point de vue éthique, quel est le risque **le plus** important que l'entreprise doive traiter ?

A) La collecte, la conservation ou la réutilisation des données au-delà des finalités justifiées
B) La complexité de l'interface rendant difficile la vérification des décisions automatisées
C) La baisse de l'exactitude du modèle lorsque les ressources sont réparties de manière inégale
D) Les mises à jour logicielles entraînant des problèmes de compatibilité entre les plateformes

37 / 40
Une librairie en ligne utilise un système d'IA pour analyser l'historique de navigation, les tendances d'achat et les précommandes afin de prévoir la demande et de s'assurer que les livres populaires soient disponibles à temps. Par la suite, l'entreprise réutilise ces mêmes données clients pour entraîner un autre modèle d'IA qui établit des profils marketing détaillés et envoie des publicités ciblées à chaque acheteur.

Quel principe de protection de la vie privée est **le plus directement** enfreint ici ?

A) Minimisation des données
B) Limitation de la conservation des données
C) Transparence et consentement
D) Limitation de l'utilisation et spécification de la finalité

38 / 40
Une organisation souhaite mettre en place un cadre pour régir l'utilisation responsable de l'IA dans l'ensemble de ses activités.

Elle utilise la norme ISO/IEC 27001 pour le management de la sécurité de l'information. L'équipe chargée de la sécurité souhaite mettre en place sa gouvernance de l'IA à l'aide d'une norme ISO/IEC qui remplisse, pour la gouvernance de l'IA, le même rôle structurel que la norme ISO/IEC 27001 pour la sécurité de l'information.

De quelle norme s'agit-il ?

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

39 / 40
Une entreprise souhaite utiliser un système d'IA pour sélectionner et classer automatiquement les candidats à un poste, sur la base de leurs CV.

Selon le règlement sur l'IA, dans quelle catégorie l'utilisation de ce système d'IA doit-elle être classée ?

A) Risque inacceptable
B) Haut risque
C) Risque limité
D) Risque minimal ou nul

40 / 40
Une équipe de développement en IA prépare des données d'entraînement pour un nouveau modèle d'apprentissage automatique. Le chef d'équipe souhaite s'assurer que l'approche de collecte des données favorise la conformité en matière de droits d'auteur.

Quelle est la pratique **la plus éthique** en matière de collecte de données ?

A) Collecter des données uniquement à partir de sources accessibles au public
B) Créer l'ensemble des données d'entraînement au sein de l'organisation
C) Télécharger des données provenant de sources dont l'autorisation est implicite
D) N'utiliser que des données largement partagées en ligne

Solutions à l'examen

1 / 40
Un responsable de la sécurité informatique travaille dans une entreprise de logistique. L'entreprise a récemment décidé d'adopter des outils d'IA pour optimiser ses itinéraires de livraison. Le responsable de la sécurité informatique dirige les efforts visant à mettre en place la sécurité de l'IA. Les deux premières étapes du cadre G.U.A.R.D. sont déjà achevées :

- L'étape « Govern » (Gouverner) a été réalisée en mettant en place une politique de sécurité de l'IA.
- L'étape « Understand » (Comprendre) a été réalisée en répertoriant tous les actifs liés à l'IA et leurs risques.

Quelle est la **prochaine** étape que le responsable de la sécurité informatique doit entreprendre ?

A) Adapter les pratiques de sécurité pour y inclure la modélisation des menaces spécifiques à l'IA, les tests et les contrôles de la chaîne d'approvisionnement
B) Analyser les risques liés à l'IA identifiés et établir une stratégie de hiérarchisation des risques en fonction de leur impact potentiel
C) Appliquer des contrôles de sécurité à l'ensemble des systèmes d'IA afin d'instaurer une sécurité de l'IA responsable tout au long de son cycle de vie
D) Évaluer la politique de sécurité de l'IA élaborée lors de l'étape « Govern » afin d'y inclure les actifs liés à l'IA répertoriés et leurs risques

A) Correct. Après les étapes « Govern » (Gouverner) et « Understand » (Comprendre), la prochaine étape du modèle G.U.A.R.D est « Adapt » (Adapter), qui consiste à mettre à jour les pratiques de sécurité afin d'y intégrer les menaces spécifiques à l'IA, les tests de sécurité et les mesures relatives à la chaîne d'approvisionnement. (Référence : A, chapitre 0)
B) Incorrect. Cela relève de l'étape « Understand » (Comprendre), qui est déjà terminée.
C) Incorrect. L'application généralisée de contrôles sans prise en compte du contexte n'est pas conforme aux étapes du cadre G.U.A.R.D. L'étape suivante consiste à adapter les pratiques, et non à appliquer des contrôles de manière aveugle. Les contrôles seront mis en œuvre à un stade ultérieur.
D) Incorrect. La révision de la politique fait partie de la phase « Govern ». L'étape suivante après « Understand » (Comprendre) est « Adapt » (Adapter), et non la refonte de la gouvernance.

2 / 40
Une entreprise examine sa stratégie en matière d'IA afin de s'assurer qu'elle profite à l'activité tout en étant éthique et solide sur le plan opérationnel. Le responsable de la sécurité de l'IA classe les deux dimensions principales de l'IA responsable et de l'IA digne de confiance.

Quelle est la **principale** différence entre l'IA responsable et l'IA digne de confiance ?

A) - L'IA responsable met l'accent sur l'éthique, la société et la gouvernance.
- L'IA digne de confiance met l'accent sur les aspects techniques et opérationnels, la transparence et l'explicabilité.

B) - L'IA responsable met l'accent sur la conformité réglementaire aux lois et législations applicables.
- L'IA digne de confiance met l'accent sur la protection des données et la cybersécurité afin de respecter les principes de protection de la vie privée.

C) - L'IA responsable met l'accent sur l'exactitude du système et les indicateurs de performance technique.
- L'IA digne de confiance met l'accent sur les lignes directrices éthiques et l'impact social afin de susciter la confiance des consommateurs.

D) - L'IA responsable met l'accent sur la sélection des fournisseurs, les risques et la gestion des coûts.
- L'IA digne de confiance met l'accent sur une formation des utilisateurs sûre et fiable ainsi que sur l'adoption des systèmes d'IA.

A) Correct. L'IA responsable met l'accent sur l'éthique, la société et la gouvernance (principes, politiques, supervision), tandis que l'IA digne de confiance met l'accent sur les aspects plus techniques et opérationnels de l'utilisation de l'IA au sein d'une organisation, tels que la robustesse, la fiabilité, la transparence et l'explicabilité. (Bibliographie : A, chapitre 0)

B) Incorrect. Cela réduit l'IA responsable à la conformité réglementaire et l'IA digne de confiance à la protection de la vie privée et à la cybersécurité. Il s'agit là de sous-ensembles qui ne rendent pas compte de la différence plus large.

C) Incorrect. Cela inverse les priorités : l'exactitude et les performances relèvent de la fiabilité technique, tandis que l'éthique et l'impact social sont au cœur de l'IA responsable.

D) Incorrect. Le choix des fournisseurs, les coûts et la formation des utilisateurs relèvent de la gestion du changement, et non de la distinction conceptuelle fondamentale entre l'IA responsable et l'IA digne de confiance.

3 / 40
Une équipe de cybersécurité dispose déjà d'un cadre de détection des menaces. Elle doit désormais également protéger un nouveau grand modèle de langage (LLM) utilisé dans le service client pour un agent conversationnel (chatbot).

L'équipe souhaite utiliser les mêmes outils que ceux qu'elle utilise pour les applications web :
- Règles de détection basées sur les signatures
- Contrôles fixes de validation des entrées

Pourquoi cette approche **n'est-elle pas** suffisante pour sécuriser le LLM ?

A) Parce que les systèmes d'IA présentent des surfaces d'attaque différentes et nécessitent des contrôles spécifiques pour le comportement du LLM et son pipeline de données
B) Parce que le cadre de détection des menaces n'inclut pas de procédures de réponse aux incidents, ce qui le rend inadapté aux LLM
C) Parce que les LLM génèrent des sorties trop variables pour être évaluées par des systèmes basés sur des règles et doivent être examinés par un humain
D) Parce que les LLM traitent les requêtes plus rapidement que les applications web, ce qui signifie que les règles basées sur les signatures ne peuvent pas suivre leur rythme
E) Parce que les pare-feu d'applications web ne sont pas sous licence pour une utilisation avec les LLM et nécessitent des contrats distincts avec les fournisseurs

A) Correct. Les outils de cybersécurité conventionnels, tels que les pare-feu d'applications web et les systèmes de détection basés sur les signatures, sont conçus pour identifier des schémas d'attaque connus visant des interfaces logicielles fixes. Les systèmes d'IA introduisent des surfaces d'attaque entièrement nouvelles : les données d'entraînement peuvent être empoisonnées, le modèle lui-même peut être inversé pour extraire des informations sensibles, et les sorties peuvent être manipulées par le biais d'entrées adversariales qui ne correspondent à aucune signature connue. Ces menaces nécessitent des contrôles spécifiques à l'IA qui agissent sur le comportement du modèle et le pipeline de données, et pas seulement sur son interface réseau. (Bibliographie : A, chapitres 0 et 1)
B) Incorrect. Les cadres de cybersécurité conventionnels incluent bel et bien des procédures de réponse aux incidents. Cette option présente de manière erronée la limite de la sécurité conventionnelle pour les systèmes d'IA, qui concerne la couverture de la surface d'attaque et non des lacunes procédurales.
C) Incorrect. Bien que les sorties des grands modèles de langage (LLM) soient très variables, les contrôles conventionnels basés sur des règles peuvent tout de même contribuer à la mise en place de défenses multicouches dans les environnements d'IA. Leur limite réside dans le fait qu'ils ne traitent pas de manière adéquate les surfaces d'attaque spécifiques à l'IA, telles que la manipulation de modèles, les entrées adversariales, l'injection de prompt et les attaques visant les données d'entraînement.
D) Incorrect. La vitesse de traitement est un critère de performance et n'est pas la raison pour laquelle la détection basée sur les signatures est insuffisante pour les systèmes d'IA. La limite ici n'est pas le temps, mais la nature de la surface d'attaque.
E) Incorrect. Il n'existe aucune restriction générale de licence empêchant l'application de pare-feu d'applications web aux points de terminaison des modèles d'IA. Cette option relève d'une considération d'achat plutôt que d'une limitation de sécurité substantielle de l'approche proposée.

4 / 40
Lorsqu'elle utilise l'IA, une entreprise dispose d'actifs spécifiques à l'IA. Chaque actif spécifique à l'IA présente également ses propres menaces clés.

Quelle combinaison **correcte** associe un actif spécifique à l'IA à sa menace clé ?

A) - Actif : journaux d'audit générés par les pare-feu réseau
- Menace : épuisement des ressources d'IA

B) - Actif : identifiants d'authentification des utilisateurs finaux
- Menace : sortie contenant une injection conventionnelle

C) - Actif : code source logiciel stocké dans un système de contrôle de version
- Menace : empoisonnement direct du modèle en phase d'exécution

D) - Actif : données d'entraînement utilisées pour construire un modèle
- Menace : divulgation de données sensibles dans la sortie

A) Incorrect. Les journaux d'audit des pare-feu réseau constituent un actif général de sécurité informatique, et non un actif lié à l'IA tel que décrit dans l'OWASP AI Exchange. L'épuisement des ressources d'IA est généralement lié à une surutilisation des ressources de calcul nécessaires aux tâches d'IA, ce qui n'a pas de lien direct avec les journaux d'audit.

B) Incorrect. Les identifiants d'authentification des utilisateurs finaux constituent un sujet de préoccupation courant en matière de sécurité informatique, et non un actif lié à l'IA tel que décrit dans l'OWASP AI Exchange. Les menaces d'injection conventionnelles, telles que l'injection SQL ou l'injection de commandes, concernent davantage la sécurité des applications que les sorties de l'IA. Cette association n'est pas pertinente.

C) Incorrect. Le code source d'un logiciel peut constituer un actif spécifique à l'IA s'il est lié à des modèles d'IA. Cependant, l'empoisonnement direct du modèle en phase d'exécution concerne la manipulation d'un modèle d'IA pendant son exécution, plutôt que son code source dans un système de contrôle de version.

D) Correct. Les données d'entraînement constituent un actif critique spécifique à l'IA. La divulgation de données dans les sorties représente une menace légitime, en particulier si des données sensibles ou propriétaires issues de l'ensemble d'entraînement apparaissent par inadvertance dans les sorties générées par l'IA. Cette combinaison identifie correctement l'actif et la menace qui y est associée dans un contexte d'IA. (Bibliographie : A, chapitre 0)

5 / 40
Un cabinet d'avocats utilise un système d'IA pour l'examen et l'analyse de documents. L'équipe de sécurité applique une approche de gestion des risques. Elle a dressé un inventaire de toutes les menaces potentielles pesant sur le système d'IA.

Quelle est la **prochaine** étape que l'équipe devrait entreprendre ?

A) Appliquer un processus de modélisation des menaces afin de traduire ces menaces en un ensemble de risques concrets et hiérarchisés.
B) Désigner un responsable des risques pour chaque menace identifiée afin de garantir une protection par conception depuis le début.
C) Déployer des contrôles de sécurité pour chaque menace afin d'atténuer les risques supplémentaires tout au long du cycle de vie du système d'IA.
D) Partager la liste complète des menaces avec les parties prenantes concernées afin de les sensibiliser et d'assurer la transparence.

A) Correct. Après avoir répertorié les menaces, l'équipe doit appliquer un processus de modélisation des menaces, qui sert de passerelle entre une liste de menaces et un ensemble de risques concrets et hiérarchisés. Cela permet à l'organisation de se concentrer sur la mise en place de contrôles pour les risques les plus significatifs. (Bibliographie : A, chapitre 0)
B) Incorrect. Désigner des responsables des risques avant d'analyser et de hiérarchiser les menaces revient à omettre une étape cruciale. Sans compréhension de l'impact concret et de la probabilité de chaque menace, l'attribution des responsabilités manque de contexte significatif.
C) Incorrect. Déployer des contrôles pour chaque menace sans hiérarchisation est inefficace et peut entraîner une mauvaise allocation des ressources. Une hiérarchisation via la modélisation des menaces est nécessaire avant de sélectionner et d'appliquer des contrôles.
D) Incorrect. Le partage d'une liste brute de menaces sans hiérarchisation n'aide pas les parties prenantes à prendre des décisions éclairées. Le processus de modélisation des menaces doit d'abord convertir cette liste en risques concrets et hiérarchisés pour que la communication soit pertinente.

6 / 40
Un ingénieur en sécurité de l'IA examine la conception du contrôle d'accès d'un système d'IA multi-agents. Chaque agent du système est doté d'un compte de service partagé. Ce compte dispose d'un accès en lecture et en écriture à la base de données de production, au système de fichiers interne et à la passerelle API (interface de programmation d'application) externe.

Les différents agents IA invoquent de manière autonome des outils et exécutent des workflows en plusieurs étapes sur l'ensemble des systèmes internes. La mise à disposition a été effectuée sur un compte partagé afin de simplifier la gestion des identifiants tout au long du pipeline des agents IA.

Pourquoi cette conception de compte partagé est-elle **particulièrement risquée** dans un contexte d'IA agentique ?

A) Parce que les agents IA génèrent des volumes d'appels API plus importants que les utilisateurs humains, ce qui augmente le risque que les contrôles de limitation de débit soient déclenchés et que le compte partagé soit suspendu
B) Parce que les agents IA n'utilisent pas d'authentification multifactorielle ; les identifiants partagés constituent donc un point de défaillance unique, sans autre couche de vérification, qu'un attaquant peut exploiter
C) Parce que les agents IA opèrent de manière autonome à travers plusieurs étapes de raisonnement au cours d'une même session ; un agent compromis peut donc enchaîner des actions sur tous les systèmes accessibles
D) Parce que les cadres réglementaires applicables aux systèmes d'IA exigent que chaque agent IA soit doté d'identifiants uniques ; les comptes partagés constituent donc une violation directe de la conformité
E) Parce que le partage de comptes de service entre plusieurs agents IA engendre une complexité opérationnelle excessive lors de la rotation obligatoire des identifiants et des activités de gestion des accès

A) Incorrect. Le volume d'appels API est un problème de performances et de disponibilité, et non un risque de sécurité spécifique à la conception des identifiants partagés. La limitation de débit s'applique à tous les types de comptes et n'explique pas pourquoi un compte partagé présente un risque disproportionné pour les agents par rapport aux utilisateurs humains partageant le même compte.
B) Incorrect. L'absence d'authentification multifactorielle pour les comptes de service constitue un problème de sécurité général qui s'applique aussi bien aux comptes partagés par des utilisateurs humains qu'aux comptes partagés par des agents. Cela n'explique pas pourquoi le risque est disproportionnellement plus élevé dans le contexte spécifique de l'IA agentique.
C) Correct. Le risque de sécurité évoqué dans la question est lié à une autonomie excessive. Les systèmes d'IA agentiques peuvent enchaîner de manière autonome des actions sur plusieurs systèmes, outils et étapes de raisonnement au cours d'une même session. Un agent compromis utilisant un compte partagé peut donc effectuer un large éventail d'actions non autorisées avant qu'une intervention ou un contrôle humain n'ait lieu. (Bibliographie : A, chapitres 0 et 1)
D) Incorrect. Les exigences réglementaires relatives à l'attribution des identifiants aux agents IA varient selon les juridictions et les secteurs et ne sont pas universellement établies avec le niveau de précision décrit. Cette réponse ne traite pas du risque de sécurité soulevé par la question.
E) Incorrect. Cette réponse traite d'un défi opérationnel, et non du risque de sécurité spécifique créé par la conception reposant sur un compte partagé dans un contexte d'IA agentique.

7 / 40
Un ingénieur en apprentissage automatique au sein d'une institution financière enquête sur deux incidents distincts impliquant le modèle de détection des fraudes de l'institution.

- **Incident 1** : un attaquant externe, n'ayant accès ni à l'architecture du modèle, ni aux données d'entraînement, ni aux paramètres, soumet des milliers de transactions spécialement conçues. Celles-ci sont affinées de manière itérative en fonction du déclenchement ou non de l'alerte de fraude.
- **Incident 2** : un chercheur utilise un modèle de substitution entraîné localement, doté d'une architecture similaire, pour générer des transactions adversariales. Celles-ci parviennent à tromper le modèle de production sans interaction préalable.

De quelles deux stratégies d'évasion s'agit-il ?

A) - Incident 1 : évasion après empoisonnement
- Incident 2 : évasion à connaissance parfaite

B) - Incident 1 : évasion à connaissance partielle
- Incident 2 : attaque par transfert

C) - Incident 1 : évasion à connaissance parfaite
- Incident 2 : évasion à connaissance nulle

D) - Incident 1 : attaque par transfert
- Incident 2 : évasion à connaissance partielle

E) - Incident 1 : évasion à connaissance nulle
- Incident 2 : attaque par transfert

A) Incorrect. L'incident 1 est identifié à tort comme une évasion après empoisonnement. Rien ne prouve que l'attaquant ait modifié les données d'entraînement. L'attaquant s'est contenté de tester de nombreuses transactions et d'apprendre à partir des réponses du modèle. L'incident 2 est identifié à tort comme une évasion à connaissance parfaite. Le chercheur n'avait pas un accès complet au modèle réel. Il a utilisé un modèle de substitution.

B) Incorrect. L'incident 1 est identifié à tort comme une évasion à connaissance partielle. L'attaquant n'avait aucun accès à la conception, aux données ou aux paramètres du modèle. Il a uniquement utilisé les sorties du modèle. L'incident 2 est correctement identifié comme une attaque par transfert, mais la réponse complète reste erronée car l'incident 1 ne constitue pas une évasion à connaissance partielle.

C) Incorrect. L'incident 1 est identifié à tort comme une attaque d'évasion à connaissance parfaite. Cette attaque ne donne pas à l'attaquant un accès complet au modèle. L'incident 2 est identifié à tort comme une attaque d'évasion à connaissance nulle. Le chercheur a utilisé un modèle de substitution pour mettre au point l'attaque ; il s'agit donc d'une attaque par transfert.

D) Incorrect. L'incident 1 est identifié à tort comme une attaque par transfert. Dans l'incident 1, l'attaquant a testé le modèle réel à plusieurs reprises et a modifié les entrées à partir des réponses. L'incident 2 est identifié à tort comme une évasion à connaissance partielle. Le chercheur a utilisé un modèle distinct ; il s'agit donc d'une attaque par transfert.

E) Correct. L'incident 1 est correctement identifié comme une évasion à connaissance nulle. L'attaquant n'avait aucun accès interne et n'a utilisé que les sorties du modèle. L'incident 2 est correctement identifié comme une attaque par transfert. Le chercheur a mis au point l'attaque sur un modèle de substitution, et celle-ci a également trompé le modèle réel. (Bibliographie : A, chapitre 2)

8 / 40
Erick est un utilisateur d'une plateforme de réseaux sociaux qui modère automatiquement les commentaires à l'aide d'un système d'IA. Il tente de faire passer un commentaire enfreignant les règles à travers la modération automatisée de la plateforme, car il souhaite stimuler l'engagement sur sa publication. Erick ne dispose d'aucune information sur l'architecture du modèle, les paramètres ou les données d'entraînement, et il ne voit qu'un message binaire lorsqu'il tente de publier son commentaire : soit « bloqué », soit « publié ». En apportant à plusieurs reprises de légères modifications à la formulation de son commentaire, Erick finit par trouver une version qui est publiée.

Quel type de menace d'évasion liée aux données d'entrée est décrit ici ?

A) Évasion à connaissance nulle
B) Évasion à connaissance partielle
C) Évasion à connaissance parfaite
D) Évasion après empoisonnement
E) Attaque par transfert

A) Correct. Erick ne dispose d'aucune information sur le modèle d'IA et utilise une stratégie d'attaque par requêtes pour trouver une entrée d'attaque réussie, lui permettant de publier son commentaire. Cela correspond à la définition d'une évasion à connaissance nulle. (Bibliographie : A, chapitre 2)
B) Incorrect. Une évasion à connaissance partielle impliquerait des informations contextuelles guidant l'attaque, dont Erick ne dispose pas.
C) Incorrect. Une évasion à connaissance parfaite nécessiterait un accès complet à l'architecture, aux paramètres et aux données d'entraînement du modèle d'IA pour élaborer des entrées avec précision.
D) Incorrect. L'évasion après empoisonnement fait référence à une évasion rendue possible par l'empoisonnement préalable des données d'entraînement du modèle d'IA. Ce n'est pas ce qui s'est produit ici.
E) Incorrect. Une attaque par transfert consisterait à créer des exemples adversariaux à l'aide d'un modèle de substitution, puis à appliquer ces exemples adversariaux au modèle cible.

9 / 40
Une société de services juridiques développe un agent conversationnel (chatbot) capable de répondre aux questions des utilisateurs et d'extraire des informations à partir de ressources internes et externes, telles que des sites web et des documents d'entreprise. Lors d'une démonstration, le comportement de l'agent conversationnel déclenche une enquête menée par l'ingénieure en sécurité de l'IA. Elle identifie deux menaces distinctes :

- **Menace 1 :** un message conçu pour passer outre les instructions et les règles de sécurité de l'agent conversationnel peut amener ce dernier à suivre les commandes du testeur plutôt que la politique que sa configuration lui impose de suivre.
- **Menace 2 :** une page web d'un fournisseur récupérée par l'agent conversationnel contient du texte intégré qui agit comme des instructions. Lorsque l'agent conversationnel ingère cette page, il traite le contenu comme des commandes et s'écarte de sa politique.

De quels types de menaces s'agit-il ?

A) - Menace 1 : évasion
- Menace 2 : empoisonnement des données

B) - Menace 1 : empoisonnement des données
- Menace 2 : injection de prompt directe

C) - Menace 1 : injection de prompt directe
- Menace 2 : injection de prompt indirecte

D) - Menace 1 : injection de prompt indirecte
- Menace 2 : évasion

A) Incorrect. La menace n° 1 n'est pas une manœuvre d'évasion, car aucune entrée adversariale n'a induit le modèle en erreur. La menace n° 2 n'est pas un empoisonnement des données, car les données d'entraînement du modèle n'ont pas été altérées.

B) Incorrect. La menace n° 1 n'est pas un empoisonnement des données, car les données d'entraînement du modèle n'ont pas été altérées. La menace n° 2 n'est pas une injection de prompt directe, car les instructions provenaient indirectement d'un contenu web.

C) Correct. La menace n° 1 est une injection de prompt directe, où les prompts de l'utilisateur trompent l'IA. La menace n° 2 est une injection de prompt indirecte, où le contenu web contient des instructions cachées. (Bibliographie : A, chapitre 2)

D) Incorrect. La menace n° 1 n'est pas une injection de prompt indirecte, car les instructions provenaient directement de l'utilisateur. La menace n° 2 n'est pas une technique d'évasion, car le modèle n'a pas été induit en erreur par des perturbations des entrées.

10 / 40
Une société de conseil dispose d'un système d'IA destiné à faciliter ses tâches quotidiennes. À l'issue d'une réunion avec un client, l'assistant IA rédige un résumé de l'entretien. L'assistant IA commence à insérer des liens internes confidentiels dans le brouillon d'e-mail destiné au client. Cependant, avant l'envoi, un détecteur d'anomalies autonome suspend l'envoi et met le brouillon en quarantaine.

L'architecte de la sécurité de l'IA souhaite limiter à l'avance l'accès de l'assistant IA chargé des réunions aux seuls droits des personnes concernées.

Quelles couches de protection contre l'injection de prompt sont illustrées ici ?

A) Contrôle automatisé et moindre privilège fondé sur l'utilisateur
B) Autorisation « juste à temps » et contrôle humain
C) Alignement du modèle et moindre privilège fondé sur l'intention
D) Gestion des entrées/sorties face à l'injection de prompt et limitation du débit

A) Correct. Un détecteur d'anomalies autonome qui suspend et met en quarantaine le brouillon relève du contrôle automatisé. Le fait de n'accorder à l'assistant que les autorisations des personnes qu'il assiste, attribuées à l'avance, relève du moindre privilège fondé sur l'utilisateur. (Bibliographie : A, chapitre 2)
B) Incorrect. Il n'y a pas de niveau d'autorisation supérieur ni d'approbation par action justifiant une autorisation « juste à temps », et le contrôle est automatisé, et non humain.
C) Incorrect. Le scénario ne mentionne pas l'alignement du modèle, et le moindre privilège décrit repose sur des droits pré-attribués en fonction de l'identité de l'utilisateur, et non sur des privilèges dynamiques liés à l'intention.
D) Incorrect. Aucune limitation de débit n'est décrite, et le contrôle présenté est un détecteur d'anomalies autonome (contrôle automatisé), et non des mesures spécifiques d'étiquetage des E/S.

11 / 40
Un auditeur évalue les risques liés à la protection de la vie privée d'un grand modèle de langage (LLM) déployé dans le secteur de la santé, ainsi que sa conformité aux lois sur la protection des données. Pour ce faire, l'auditeur interroge le modèle à l'aide de requêtes ciblées concernant un patient spécifique et des non-patients synthétiques soigneusement appariés. En observant les différences dans les schémas de réponse et les scores de confiance du modèle, l'auditeur peut déterminer de manière fiable si le dossier de ce patient a été inclus dans les données d'entraînement, même si aucun détail médical réel n'est révélé.

Quel type d'attaque est simulé ici ?

A) Divulgation de données dans la sortie du modèle
B) Injection de prompt indirecte
C) Inférence d'appartenance
D) Inversion de modèle

A) Incorrect. La divulgation de données dans la sortie du modèle se produit lorsque le modèle révèle directement des informations sensibles contenues dans les données d'entraînement ou dans le contexte de la phase d'exécution. Dans ce scénario, aucune information médicale réelle n'est divulguée.
B) Incorrect. L'injection de prompt indirecte se produit lorsqu'un tiers intègre des instructions adversariales dans un contenu qui est récupéré et inséré dans un prompt, manipulant ainsi le comportement du modèle. Le scénario ne décrit pas une telle injection.
C) Correct. L'inférence d'appartenance se produit lorsqu'un attaquant présente au modèle une entrée identifiant un individu spécifique et utilise les signaux de confiance du modèle pour déterminer si cet individu figure dans l'ensemble d'entraînement. (Bibliographie : A, chapitre 2)
D) Incorrect. L'inversion de modèle se produit lorsqu'un attaquant reconstitue une partie de l'ensemble d'entraînement par des expérimentations intensives, en optimisant les entrées afin de maximiser les signaux de confiance dans la sortie du modèle. Ce type d'attaque reconstitue des attributs ou des enregistrements sensibles à partir du modèle, et ne se limite pas à déterminer si un enregistrement figurait dans l'ensemble d'entraînement.

12 / 40
Quelle est l'une des **principales** conséquences d'une attaque d'exfiltration du modèle réussie ?

A) L'attaquant peut extraire l'ensemble de données d'entraînement brut utilisé pour construire le modèle d'origine, qui peut potentiellement contenir des enregistrements sensibles.
B) L'attaquant prend le contrôle direct du modèle déployé et peut modifier ses poids et ses paramètres sans autorisation préalable.
C) L'attaquant obtient une réplique du modèle qui peut être utilisée pour orchestrer des attaques d'évasion sans déclencher les défenses de l'original.
D) Le modèle d'origine devient indisponible pour les utilisateurs légitimes en raison de l'épuisement des ressources, ce qui entraîne une dégradation du service et des délais d'attente répétés.

A) Incorrect. L'exfiltration du modèle reproduit le comportement du modèle à travers des paires d'entrée-sortie, mais elle n'extrait pas l'ensemble de données d'entraînement d'origine. L'extraction des données d'entraînement constitue une menace distincte (inversion de modèle ou inférence d'appartenance).
B) Incorrect. L'exfiltration du modèle ne permet pas de modifier le modèle déployé d'origine. Elle aboutit à une copie des fonctionnalités du modèle, et non à un contrôle sur l'original.
C) Correct. L'une des principales conséquences de l'exfiltration du modèle est que l'attaquant peut utiliser la réplique du modèle pour mener des attaques d'évasion, en échappant à la limitation de débit, au contrôle d'accès et aux mécanismes de détection présents dans le système d'origine. (Bibliographie : A, chapitre 2)
D) Incorrect. Le déni de service (DoS) est une conséquence de l'épuisement des ressources, qui constitue une menace distincte. L'exfiltration du modèle n'a pas d'impact direct sur la disponibilité du modèle d'origine.

13 / 40
Un grand modèle de langage (LLM) hébergé dans le cloud est victime d'une attaque éponge. Les attaquants envoient des entrées très longues ou complexes qui imposent un surcroît de calcul, ce qui fait grimper les coûts et provoque des ralentissements ou des pannes. Afin de remédier à cette situation, l'ingénieur en sécurité de l'IA souhaite mettre en place deux contrôles :

- **Le contrôle 1** validera et assainira les entrées afin de rejeter ou de corriger les contenus malveillants, tels que les entrées anormalement volumineuses.
- **Le contrôle 2** limitera les ressources de calcul consommées par entrée du modèle afin d'empêcher toute surutilisation.

Quels sont ces deux contrôles ?

A) - Contrôle 1 : gestion des entrées anormales
- Contrôle 2 : masquage des limites de confiance

B) - Contrôle 1 : validation des entrées contre le déni de service (DoS)
- Contrôle 2 : limitation des ressources

C) - Contrôle 1 : contrôle d'accès au modèle
- Contrôle 2 : modèle résistant à l'évasion

D) - Contrôle 1 : surveillance de l'utilisation
- Contrôle 2 : limitation de débit

A) Incorrect. La gestion des entrées anormales détecte les entrées qui s'écartent de la distribution des données d'entraînement et concerne principalement les menaces d'évasion et d'exfiltration du modèle. Le masquage des limites de confiance restreint les informations de sortie afin de réduire la capacité de l'attaquant à optimiser les échantillons adversariaux.

B) Correct. Il s'agit des deux contrôles présentés dans la question et spécifiquement conçus pour lutter contre l'épuisement des ressources induit par les entrées, en complément des contrôles généraux. (Bibliographie : A, chapitre 2)

C) Incorrect. Le contrôle d'accès au modèle restreint les personnes autorisées à interagir avec le modèle et constitue un contrôle général applicable à de multiples menaces liées aux entrées. Le modèle résistant à l'évasion est un contrôle mis en place lors de la phase de développement, conçu pour accroître la résilience du modèle face aux attaques par évasion.

D) Incorrect. La surveillance consiste à observer, corréler et consigner l'utilisation du modèle, les entrées, les sorties et le comportement du système afin d'identifier des événements ou des schémas susceptibles d'indiquer un incident de (cyber)sécurité. La limitation de débit consiste à limiter le débit (la fréquence) d'accès au modèle.

14 / 40
Un ingénieur en sécurité de l'IA identifie les menaces suivantes pour une entreprise :

- **Des prompts adversariaux** sont conçus par des attaquants pour provoquer des erreurs de classification.
- Un **pipeline d'ingestion compromis** introduit des enregistrements contenant des portes dérobées dans le corpus avant la construction du modèle d'IA.
- Les attaquants altèrent les **workflows d'étiquetage** en modifiant les annotations de référence afin de créer un biais dans le comportement du modèle d'IA.
- Les **données d'entraînement** proviennent de sources externes ou tierces sans validation appropriée.

Selon l'OWASP AI Exchange, quelle menace **n'implique pas** de risques d'empoisonnement des données (phase de développement) ?

A) Entrées adversariales
B) Pipeline d'ingestion compromis
C) Workflows d'étiquetage
D) Données d'entraînement

A) Correct. Cela décrit des exemples adversariaux au moment du déploiement, et non l'empoisonnement des données (phase de développement). (Bibliographie : A, chapitre 3)
B) Incorrect. L'introduction d'éléments truqués dans le corpus collecté avant la construction du modèle crée une porte dérobée pendant le développement, ce qui constitue un risque d'empoisonnement en phase de développement.
C) Incorrect. L'altération des étiquettes lors de l'annotation empoisonne l'ensemble de données avant la construction du modèle, ce qui constitue une attaque au cours de la phase de développement.
D) Incorrect. L'utilisation d'ensembles de données externes non vérifiés introduit le risque que des échantillons empoisonnés soient intégrés dans les sources de données. L'OWASP AI Exchange met en évidence les risques liés à la chaîne d'approvisionnement et la nécessité d'un contrôle de la qualité des données.

15 / 40
Selon l'OWASP AI Exchange, quel exemple illustre un **empoisonnement direct du modèle en phase de développement** ?

A) Les attaquants créent des entrées adversariales en phase d'exécution pour induire le modèle en erreur
B) Les attaquants manipulent les poids et le pipeline d'entraînement pour modifier le comportement du modèle.
C) Des données d'entraînement sensibles sont involontairement exposées via les sorties du modèle.
D) Un pic de trafic submerge le service d'IA, entraînant un ralentissement ou une interruption de service.

A) Incorrect. Il s'agit d'une attaque adversariale en phase d'exécution, et non d'un empoisonnement en phase de développement.
B) Correct. L'empoisonnement direct du modèle en phase de développement implique de compromettre l'environnement de développement, le pipeline d'entraînement ou les paramètres du modèle. L'attaquant peut injecter de la logique, modifier les poids ou altérer les procédures d'entraînement afin d'y intégrer un comportement malveillant. (Bibliographie : A, chapitre 3)
C) Incorrect. Cela fait référence à des menaces pour la confidentialité telles que l'inversion de modèle ou la divulgation de données, et non à l'empoisonnement de modèle.
D) Incorrect. Il s'agit d'un problème de déni de service (DoS), et non d'empoisonnement de modèle.

16 / 40
Selon l'OWASP AI Exchange, quel exemple illustre un **empoisonnement de modèle dans la chaîne d'approvisionnement** ?

A) Les attaquants compromettent les données, les modèles ou les composants de la chaîne d'approvisionnement avant leur intégration.
B) Les attaquants manipulent les données d'entrée lors de la phase d'exécution afin d'induire le modèle en erreur et de lui faire produire des sorties erronées.
C) Le modèle produit des résultats erronés en raison d'une qualité insuffisante des données ou de l'entraînement.
D) Le modèle divulgue involontairement des données d'entraînement sensibles par le biais de ses réponses.

A) Correct. L'attaque se produit avant le déploiement du modèle d'IA, via des sources tierces telles que des modèles pré-entraînés, des ensembles de données externes et des composants ou pipelines d'apprentissage automatique (ML). L'OWASP AI Exchange souligne la nécessité d'une gestion de la chaîne d'approvisionnement pour garantir la confiance et l'intégrité. (Bibliographie : A, chapitre 3)
B) Incorrect. Cela décrit des attaques d'évasion ou d'injection de prompt, et non l'empoisonnement de modèle dans la chaîne d'approvisionnement. Ces attaques se produisent pendant l'utilisation du modèle, et non pendant son développement ou son acquisition.
C) Incorrect. Les résultats erronés dus à un manque de données ou à une qualité insuffisante de l'entraînement ne constituent pas des attaques. L'empoisonnement de la chaîne d'approvisionnement implique spécifiquement une altération malveillante intentionnelle.
D) Incorrect. Cela fait référence à des menaces pour la confidentialité telles que l'inversion de modèle ou la divulgation de données sensibles, et non à l'empoisonnement de modèle.

17 / 40
Une équipe de développement en IA procède à une évaluation des risques de son environnement de développement. Elle identifie deux menaces distinctes :

La menace 1 implique un accès non autorisé aux ensembles de données utilisés pour entraîner et tester son modèle d'IA.
La menace 2 implique un accès non autorisé aux paramètres et aux poids du modèle stockés dans un référentiel compromis.

De quels types de menaces s'agit-il ?

A) - Menace 1 : fuite de données en phase de développement
- Menace 2 : fuite directe du modèle en phase de développement

B) - Menace 1 : fuite directe des données d'augmentation
- Menace 2 : fuite du référentiel

C) - Menace 1 : fuite directe de données en phase d'exécution
- Menace 2 : fuite directe du modèle en phase d'exécution

D) - Menace 1 : fuite des données d'entrée
- Menace 2 : fuite du code source ou de la configuration

A) Correct. Une fuite de données en phase de développement implique un accès non autorisé aux données d'entraînement et de test, portant atteinte à la confidentialité des données sensibles ou de la propriété intellectuelle contenues dans ces ensembles de données. Une fuite directe du modèle en phase de développement implique un accès non autorisé aux attributs du modèle (paramètres, poids, architecture du modèle), portant atteinte à la confidentialité de la propriété intellectuelle du modèle. (Bibliographie : A, chapitre 3)

B) Incorrect. La fuite directe des données d'augmentation ne correspond pas aux ensembles de données d'entraînement et de test, et la notion de « fuite du référentiel » est trop vague. Le problème concerne spécifiquement une fuite d'artefacts du modèle, et non pas simplement un problème générique lié au référentiel.

C) Incorrect. Les menaces décrites ne sont pas des fuites en phase d'exécution. Elles se produisent dans l'environnement de développement.

D) Incorrect. La fuite des données d'entrée fait généralement référence à la fuite d'entrées fournies par l'utilisateur, et non aux ensembles de données d'entraînement et de test. Les poids du modèle ne constituent pas du code source ou une configuration, mais des artefacts du modèle.

18 / 40
Une équipe de sécurité examine le paysage des menaces de sécurité en phase d'exécution d'un système d'IA. Ce système d'IA est intégré à une application web. L'équipe identifie deux menaces de sécurité conventionnelles :

- **Menace 1 :** un attaquant peut accéder à la mémoire du serveur de production et extraire les poids et les paramètres du modèle.
- **Menace 2 :** un attaquant ayant accès au registre des modèles remplace l'artefact du modèle déployé par une version contenant un cheval de Troie.

Quelles sont les conséquences de ces menaces de sécurité conventionnelles qui sont **spécifiques** aux systèmes d'IA ?

A) - La menace 1 entraîne le réentraînement automatique du système sur des données malveillantes.
- La menace 2 force le modèle à rejeter des prompts valides.

B) - La menace 1 modifie l'architecture du modèle en phase d'exécution.
- La menace 2 compresse silencieusement les poids pour réduire leur précision.

C) - La menace 1 désactive les filtres de sécurité, rendant le système plus permissif de par sa conception.
- La menace 2 rend les sorties déterministes et prévisibles.

D) - La menace 1 permet le vol de modèles et l'inférence potentielle des données d'entraînement.
- La menace 2 permet l'altération non détectée des modèles.

A) Incorrect. La menace 1 ne concerne pas l'empoisonnement par réentraînement automatique, mais le vol de modèles et l'inférence de données. La menace 2 concerne l'altération de l'intégrité et les portes dérobées, et non le rejet de prompts.

B) Incorrect. Les modifications de l'architecture en phase d'exécution et la compression silencieuse de la précision ne sont pas des conséquences spécifiques à l'IA de ces menaces. Elles décrivent un comportement peu plausible ou non malveillant, plutôt qu'un vol ou une altération dissimulée.

C) Incorrect. La désactivation des filtres de sécurité correspond à une altération, et non au risque d'exfiltration de la menace 1. Rendre les sorties déterministes ne tient pas compte de la nature liée à l'intégrité de la menace 2.

D) Correct. La menace 1 conduit au vol de modèles et à une inférence potentielle des données d'entraînement. La menace 2 conduit à des attaques dissimulées contre l'intégrité (altération non détectée du modèle). (Bibliographie : A, chapitre 4)

19 / 40
Une entreprise de services financiers déploie un modèle d'apprentissage automatique (ML) pour détecter les transactions frauduleuses. Le modèle ML est hébergé dans son environnement de production et traite en continu des données de transactions en temps réel. Deux incidents de sécurité se produisent :

- **Incident 1** : un attaquant obtient un accès non autorisé à l'environnement du modèle déployé et modifie les paramètres du modèle de sorte que certaines transactions frauduleuses ne soient plus signalées. Cette manipulation n'est pas détectée, et l'attaquant en tire profit.
- **Incident 2** : un attaquant exploite une vulnérabilité du système de production. Cela lui permet de répliquer le modèle et d'analyser son comportement en vue d'autres attaques.

D'après l'OWASP AI Exchange, de quel type d'incidents de sécurité s'agit-il ?

A) - Incident A : empoisonnement des données
- Incident B : empoisonnement direct du modèle en phase d'exécution

B) - Incident A : empoisonnement des données
- Incident B : attaque d'évasion

C) - Incident A : empoisonnement direct du modèle en phase d'exécution
- Incident B : fuite directe du modèle en phase d'exécution

D) - Incident A : fuite directe du modèle en phase d'exécution
- Incident B : attaque d'évasion

A) Incorrect. La fuite directe du modèle en phase d'exécution identifie correctement l'incident de sécurité B. L'attaquant extrait les paramètres et l'architecture du modèle, compromettant ainsi la confidentialité de celui-ci. Cela permet sa reproduction et une exploitation ultérieure. Cependant, l'empoisonnement des données se produit pendant la phase de développement via la manipulation des données d'entraînement. Le scénario décrit une compromission en phase d'exécution, et non une altération pendant la phase d'entraînement.

B) Incorrect. L'empoisonnement des données se produit pendant la phase de développement via la manipulation des données d'entraînement. Le scénario décrit une compromission en phase d'exécution, et non une altération pendant la phase d'entraînement. Une attaque d'évasion implique la manipulation des entrées, et non la modification du modèle lui-même. Dans le scénario, le modèle est directement altéré.

C) Correct. L'empoisonnement direct du modèle en phase d'exécution identifie correctement l'incident de sécurité A. L'attaquant modifie le modèle déployé, altérant ainsi son comportement. Il s'agit d'une attaque contre l'intégrité du modèle connue sous le nom d'empoisonnement en phase d'exécution (reprogrammation). La fuite directe du modèle en phase d'exécution identifie correctement l'incident de sécurité B. L'attaquant extrait les paramètres et l'architecture du modèle, compromettant ainsi la confidentialité de celui-ci. (Bibliographie : A, chapitre 4)

D) Incorrect. La fuite directe du modèle en phase d'exécution identifie correctement l'incident de sécurité B. L'attaquant extrait les paramètres et l'architecture du modèle, compromettant ainsi la confidentialité de celui-ci. Cela permet sa reproduction et une exploitation ultérieure. Cependant, une attaque d'évasion consiste à manipuler les entrées, et non à modifier le modèle lui-même. Dans le scénario, le modèle est directement altéré. (Bibliographie : A, chapitre 4)

20 / 40
Un établissement de santé déploie un assistant IA pour aider les cliniciens à synthétiser les dossiers des patients et à formuler des recommandations. Le système s'intègre à des bases de données internes et à des outils externes, et les cliniciens s'appuient sur les sorties qu'il génère dans le cadre de leurs activités quotidiennes. Deux menaces sont observées :

- **Menace 1** : l'assistant IA génère une réponse contenant un script malveillant caché intégré dans du contenu HTML. Lorsque la sortie est consultée, le script s'exécute et envoie les données de session à l'attaquant.
- **Menace 2** : un attaquant exploite des contrôles d'accès insuffisants dans le système pour intercepter et récupérer des données sensibles sur les patients.

D'après l'OWASP AI Exchange, quels types de menaces sont décrits ?

A) - Menace 1 : empoisonnement des données
- Menace 2 : fuite des données d'entrée

B) - Menace 1 : empoisonnement des données
- Menace 2 : attaque par inversion de modèle

C) - Menace 1 : sortie contenant une injection conventionnelle
- Menace 2 : fuite des données d'entrée

D) - Menace 1 : sortie contenant une injection conventionnelle
- Menace 2 : attaque par inversion de modèle

A) Incorrect. La menace 2 est correctement identifiée comme une fuite des données d'entrée. Les données sensibles fournies au modèle sont exposées ou interceptées, ce qui constitue une violation de la confidentialité. Cependant, l'empoisonnement des données se produit pendant la phase de développement, par la modification des données d'entraînement. Le scénario décrit une injection conventionnelle en phase d'exécution, et non une manipulation des données d'entraînement.

B) Incorrect. L'empoisonnement des données se produit pendant la phase de développement, par la modification des données d'entraînement. Le scénario décrit une injection conventionnelle en phase d'exécution, et non une manipulation des données d'entraînement. L'inversion de modèle consiste à déduire les données d'entraînement à partir des sorties. Le scénario ne décrit pas une situation dans laquelle une inférence de modèle a eu lieu.

C) Correct. La menace 1 est correctement identifiée comme une sortie contenant une injection conventionnelle. La sortie de l'IA comprend du contenu exécutable malveillant, en l'occurrence une injection de script. La menace 2 est correctement identifiée comme une fuite des données d'entrée. Les données sensibles fournies au modèle sont exposées ou interceptées, ce qui constitue une violation de la confidentialité. (Bibliographie : A, chapitre 4)

D) Incorrect. La menace 1 est correctement identifiée comme une sortie contenant une injection conventionnelle. La sortie de l'IA comprend du contenu exécutable malveillant, en l'occurrence une injection de script. Cependant, l'inversion de modèle consiste à déduire les données d'entraînement à partir des sorties. Le scénario ne décrit pas une situation dans laquelle une inférence de modèle a eu lieu.

21 / 40
Une entreprise utilise un système de génération augmentée par récupération (RAG) pour répondre aux questions de ses employés concernant les politiques internes. Le système d'IA extrait des documents d'une base de données vectorielle et les ajoute au prompt avant de les envoyer au modèle d'IA.

Un audit de sécurité révèle qu'un attaquant a obtenu un accès non autorisé à la base de données vectorielle. L'attaquant a consulté les documents d'augmentation stockés et a remplacé un document de politique légitime par un autre, contenant de fausses informations. Les employés qui interrogent le système d'IA au sujet de cette politique reçoivent désormais des conseils erronés.

Quelles menaces de sécurité ce scénario illustre-t-il ?

A) - Menace 1 : fuite des données d'augmentation
- Menace 2 : sortie contenant une injection conventionnelle

B) - Menace 1 : manipulation des données d'augmentation
- Menace 2 : fuite des données d'augmentation

C) - Menace 1 : inférence d'appartenance
- Menace 2 : manipulation des données d'augmentation

D) - Menace 1 : sortie contenant une injection conventionnelle
- Menace 2 : inférence d'appartenance

A) Incorrect. Une sortie contenant une injection conventionnelle implique un contenu exécutable malveillant intégré à la sortie du modèle, ce qui n'est pas le cas dans ce scénario. Bien que la fuite des données d'augmentation soit correctement identifiée, la réponse complète est incorrecte.

B) Correct. La manipulation des données d'augmentation se produit lorsqu'un attaquant modifie les données d'augmentation stockées dans une base de données vectorielle, en y insérant de fausses informations qui altèrent le comportement du modèle. La fuite des données d'augmentation survient parce qu'un attaquant a obtenu un accès non autorisé aux données d'augmentation stockées dans la base de données vectorielle. La base de données vectorielle représente un emplacement supplémentaire où les données d'augmentation peuvent être consultées ou modifiées, créant ainsi une nouvelle surface d'attaque et étendant les risques d'exposition au-delà du référentiel de documents d'origine. (Bibliographie : A, chapitre 4)

C) Incorrect. L'inférence d'appartenance est une attaque par laquelle un adversaire détermine si des données spécifiques apparaissent dans l'ensemble d'entraînement en analysant des signaux de confiance. Cela n'est pas décrit dans le scénario. Bien que la manipulation des données d'augmentation soit correctement identifiée, l'association avec l'inférence d'appartenance est incorrecte.

D) Incorrect. Une sortie contenant une injection conventionnelle fait référence à un code malveillant intégré dans la sortie du modèle. L'inférence d'appartenance est une attaque dans laquelle un adversaire détermine si des données spécifiques apparaissent dans l'ensemble d'entraînement en analysant des signaux de confiance. Aucune de ces menaces de sécurité n'est décrite dans le scénario.

22 / 40
Des contrôles de gouvernance généraux doivent être mis en œuvre pour garantir une supervision efficace de la sécurité des systèmes d'IA.

Quelle approche l'illustre le **mieux** ?

A) Définir des politiques, des rôles et des pratiques de gestion des risques clairs pour guider le développement, le déploiement et la surveillance sécurisés des systèmes d'IA
B) Intégrer des contrôles de sécurité dans les systèmes d'IA afin d'appliquer les principes de « sécurité dès la conception » lors de la phase de conception du système d'IA
C) Respecter les exigences minimales de sécurité, car les systèmes d'IA sont plus sûrs par défaut et ne nécessitent pas de multiples couches de sécurité
D) Séparer les contrôles de gouvernance de l'IA des contrôles de sécurité de l'information afin de s'assurer que ces deux concepts distincts soient traités de manière équivalente

A) Correct. Les contrôles de gouvernance consistent à définir et à faire respecter les politiques, les rôles, la responsabilité et la gestion des risques au sein d'un système de management de la sécurité de l'information (SMSI), garantissant ainsi une supervision de bout en bout tout au long du cycle de vie de l'IA. (Bibliographie : A, chapitre 1)
B) Incorrect. L'application des principes de « sécurité dès la conception » lors de la phase de conception des systèmes d'IA constitue un élément important des contrôles de gouvernance généraux, mais elle n'est pas suffisante. La gouvernance englobe les politiques, la responsabilité, les risques, la surveillance et l'amélioration continue.
C) Incorrect. Les systèmes d'IA ne sont pas sécurisés par défaut. Se contenter d'exigences minimales revient à négliger la défense en profondeur et la supervision continue nécessaires face à l'évolution des risques liés à l'IA.
D) Incorrect. Séparer la gouvernance de l'IA de la sécurité de l'information crée des cloisonnements et des lacunes. Une bonne pratique générale consiste à intégrer ces éléments aux cadres de gestion des risques d'entreprise et au système de management de la sécurité de l'information (SMSI) afin de maintenir la cohérence des contrôles et de la responsabilité.

23 / 40
Une agence de marketing de taille moyenne dispose de peu d'expérience en matière d'IA et n'a pas encore mis en place de gouvernance formelle de l'IA. Elle souhaite utiliser un outil d'IA générative pour rédiger des e-mails destinés à ses clients. Cet outil nécessite d'accéder aux noms et adresses e-mail des clients. L'entreprise devrait mettre en place une gouvernance de l'IA et établir une supervision de la sécurité avant de pouvoir utiliser cet outil d'IA.

Quel est le **strict minimum** que l'entreprise devrait faire en matière de supervision de la sécurité ?

A) Réaliser des exercices de red teaming sur le système d'e-mails basé sur l'IA générative afin d'atténuer les risques de sécurité avant son lancement
B) Recruter un spécialiste de la gouvernance de l'IA pour réaliser une évaluation des menaces et définir des politiques et des contrôles pour le déploiement
C) Dresser un inventaire des utilisations prévues de l'IA et réaliser une analyse des risques afin d'identifier les menaces, les contrôles et les responsabilités
D) Solliciter l'avis juridique d'un cabinet spécialisé dans l'IA et les exigences du règlement sur l'IA qui s'y rapportent
E) Commencer par un projet pilote et surveiller les performances, les utilisations abusives et les incidents du système de messagerie basé sur l'IA à l'aide de journaux

A) Incorrect. Le red teaming n'est pas une mesure de gouvernance minimale. Il s'agit d'une activité d'assurance intervenant à un stade ultérieur, qui repose sur un périmètre défini, des contrôles de référence et une responsabilité clairement établie par la gouvernance initiale.
B) Incorrect. Le recrutement d'un spécialiste de la gouvernance de l'IA relève d'une décision en matière de ressources, et non d'une mesure minimale de gouvernance.
C) Correct. Cette option représente la mesure de gouvernance minimale pour la supervision de la sécurité de l'IA, car elle permet d'identifier les risques, les menaces, les contrôles et les responsabilités avant le déploiement. (Bibliographie : A, chapitre 1)
D) Incorrect. L'examen juridique est important, mais il relève de la conformité et ne peut se substituer à la gouvernance de la sécurité. Retarder la mise en place de la gouvernance revient à laisser les données à caractère personnel et les risques d'utilisation abusive sans protection.
E) Incorrect. Les phases de pilotage et de surveillance sont de nature opérationnelle et réactive, et le pilotage sur un petit segment de clientèle relève d'une stratégie de mise en œuvre. En l'absence de gouvernance préalable, le pilotage et la surveillance ne disposent pas de garde-fous, de rôles et de contrôles définis pour protéger les données CRM.

24 / 40
Quelle est la **meilleure** description de la portée des contrôles de gouvernance généraux en matière de sécurité de l'IA ?

A) Un ensemble de méthodes de chiffrement utilisées pour sécuriser les sorties de l'IA après leur génération
B) Un ensemble de contrôles globaux s'appliquant à toutes les menaces liées à l'IA et à toutes les étapes de son cycle de vie
C) Des contrôles limités à la phase d'entraînement du modèle, axés sur la qualité des ensembles de données et les techniques de réglage
D) Des mesures facultatives que les organisations peuvent adopter si les systèmes d'IA ne traitent pas de données sensibles
E) Des mesures de protection techniques appliquées exclusivement en phase d'exécution pour détecter les entrées malveillantes

A) Incorrect. Le chiffrement n'est qu'un contrôle technique parmi d'autres et ne constitue pas à lui seul une gouvernance. Les contrôles généraux ne se limitent pas à la cryptographie ou aux sorties.
B) Correct. La gouvernance est transversale : elle englobe les politiques, les rôles, la responsabilité et la gestion des risques qui s'appliquent à l'ensemble des menaces et à toutes les étapes du cycle de vie. (Bibliographie : A, chapitre 1)
C) Incorrect. La gouvernance s'applique à l'ensemble du cycle de vie, et pas seulement à la phase d'entraînement.
D) Incorrect. Les contrôles de gouvernance ne sont pas facultatifs et ne se limitent pas au traitement des données sensibles.
E) Incorrect. Les contrôles généraux ne se limitent pas à la phase d'exécution ou aux défenses techniques. Ils couvrent les niveaux organisationnel, procédural et de gouvernance.

25 / 40
Une entreprise du secteur de la santé décide d'utiliser un modèle d'IA prêt à l'emploi, fourni par un tiers, pour une application. L'application traite les données saisies par l'utilisateur et récupère des données externes.

L'ingénieur en sécurité de l'IA s'inquiète de la sécurité du modèle. La stratégie de sécurité définissant la répartition des responsabilités en matière de sécurité entre le fournisseur de modèle d'IA et l'entreprise du secteur de la santé n'est pas établie.

Qu'est-ce qui est **correct** concernant ces responsabilités ?

A) La plupart des contrôles de sécurité peuvent être délégués au fournisseur lors de l'utilisation de modèles d'IA prêts à l'emploi, car l'entraînement et l'hébergement assurés par le fournisseur couvrent les risques de sécurité.
B) Le déployeur est responsable de tous les contrôles de sécurité lorsqu'il intègre le modèle d'IA dans ses systèmes et tient compte de la sécurité avant le déploiement.
C) Le fournisseur a entraîné le modèle d'IA ; il gère donc les contrôles au niveau du modèle, tandis que le déployeur doit gérer les contrôles au niveau de l'application.
D) Le fournisseur est responsable de tous les contrôles de sécurité car il a entraîné le modèle d'IA et l'héberge ; par conséquent, tous les risques et toutes les attaques devraient être gérés par le fournisseur.

A) Incorrect. L'injection de prompt et les fuites de données dans le contexte d'outils externes ou de la génération augmentée par récupération (RAG) constituent principalement des risques au niveau de la couche applicative et ne peuvent être entièrement délégués au fournisseur. L'hébergement et l'entraînement ne suffisent pas à eux seuls à remédier à ces risques.
B) Incorrect. Le fournisseur entraîne et affine le modèle et définit les contrôles de sécurité de base. Il est donc responsable d'une partie de la sécurité, ce qui implique une responsabilité partagée.
C) Correct. Le fournisseur entraîne et affine le modèle, et est responsable des contrôles au niveau du modèle. Le déployeur doit néanmoins rester vigilant face aux menaces lorsqu'il utilise des modèles prêts à l'emploi, et est responsable des contrôles au niveau de l'application. (Bibliographie : A, chapitre 0)
D) Incorrect. Bien que le fournisseur entraîne et héberge le modèle, il n'est pas responsable de l'ensemble des contrôles de sécurité. Il est responsable d'une partie de la sécurité, ce qui implique une responsabilité partagée.

26 / 40
Une organisation a déployé une application d'IA destinée aux clients utilisant un grand modèle de langage (LLM) fourni par un fournisseur tiers. L'accès au modèle s'effectue via un service d'API (interface de programmation d'application) géré et celui-ci n'a pas fait l'objet d'un entraînement supplémentaire ni d'une modification de la part de l'organisation.

L'équipe de sécurité découvre que des entrées utilisateur soigneusement élaborées peuvent amener le modèle à ignorer les limites comportementales qui lui ont été configurées et à produire des sorties qui enfreignent les restrictions de contenu de l'organisation. L'équipe examine quel contrôle mettre en œuvre qui relève de la responsabilité de l'organisation en tant que déployeur plutôt que de celle du fournisseur tiers.

Quel contrôle l'équipe de sécurité devrait-elle mettre en œuvre pour résoudre ce problème ?

A) Insérer une couche de validation des sorties qui analyse toutes les réponses du modèle avant qu'elles n'atteignent les utilisateurs finaux et bloque les sorties qui enfreignent la politique de contenu de l'organisation
B) Modifier le pipeline d'inférence du fournisseur afin d'imposer des contraintes de génération plus strictes qui limitent la génération de réponses et empêchent l'apparition de contenus non autorisés
C) Demander au fournisseur de réentraîner et d'ajuster le modèle fourni en matière de sécurité à l'aide de données supplémentaires de refus et d'alignement sur la politique, afin d'améliorer la résistance face aux entrées visant à contourner les restrictions (jailbreak)
D) Exiger du fournisseur qu'il révise et applique des règles de modération de contenu renforcées à l'échelle du service, ainsi que des mesures d'application, afin que les sorties dangereuses soient filtrées globalement avant leur diffusion

A) Correct. Les déployeurs sont responsables des contrôles mis en œuvre autour de l'application d'IA et de ses sorties, tandis que les fournisseurs tiers gèrent l'entraînement du modèle sous-jacent et le comportement d'inférence. Une couche de validation des sorties est un contrôle côté déployeur qui peut bloquer les réponses enfreignant la politique avant qu'elles n'atteignent les utilisateurs finaux. (Bibliographie : A, chapitre 0)
B) Incorrect. La modification du pipeline d'inférence du fournisseur ne relève pas du contrôle du déployeur lorsqu'il utilise une API gérée par un tiers. Au mieux, il peut ajuster les paramètres exposés, mais pas modifier le pipeline.
C) Incorrect. Le réentraînement ou le réglage de sécurité du modèle relève de la responsabilité du fournisseur. Le déployeur peut en faire la demande, mais ne peut pas le mettre en œuvre lui-même pour un modèle d'API gérée.
D) Incorrect. Les politiques de modération globales et les contrôles de filtrage à l'échelle du fournisseur sont gérés par le fournisseur tiers, et non par l'organisation qui déployeur le modèle.

27 / 40
Une équipe prépare des données d'entraînement pour un assistant d'IA générative. Elle constate qu'elle n'a pas besoin de tous les champs et de toutes les données qu'elle utilise actuellement pour garantir les performances du modèle d'IA. Dans un deuxième temps, elle envisage des contrôles visant à limiter les données sensibles et à améliorer la confidentialité et l'intégrité.

Quelle est la **meilleure** façon d'améliorer la confidentialité et l'intégrité ?

A) Conserver tous les champs dans les données d'entraînement et chiffrer les données au repos et dans les sauvegardes
B) Conserver les identifiants dans les données d'entraînement afin de simplifier les futures tâches et processus de réentraînement
C) Supprimer les champs et les données inutiles à l'entraînement et aux performances du modèle
D) N'utiliser que des données synthétiques pour toutes les phases d'entraînement et d'évaluation de l'IA générative

A) Incorrect. Le chiffrement au repos protège les données stockées, mais il ne réduit pas l'exposition inutile des données sensibles ni ne minimise la surface d'attaque. Il améliore la confidentialité uniquement au repos et n'apporte que peu de protection en matière d'intégrité.
B) Incorrect. Conserver les identifiants dans les données d'entraînement augmente le risque d'exposition. Cela compromet la confidentialité et peut nuire à l'intégrité.
C) Correct. La minimisation des données supprime les champs et les enregistrements superflus, réduisant ainsi la surface d'attaque des données. Elle renforce l'intégrité en offrant moins d'occasions aux attaquants d'effectuer des altérations sur les données et en empêchant le modèle de s'appuyer sur des schémas non pertinents. (Bibliographie : A, chapitre 1)
D) Incorrect. L'utilisation de données synthétiques n'est pas toujours faisable ni nécessaire, et elle ne garantit pas une amélioration de la confidentialité et de l'intégrité. Les données synthétiques peuvent tout de même conserver des schémas sensibles ou introduire d'autres risques.

28 / 40
Une équipe produit entraîne un modèle d'IA destiné au service client. L'ensemble de données d'entraînement actuel comprend les adresses complètes des clients, leurs numéros de téléphone et des enregistrements hérités qui ne sont pas nécessaires pour atteindre les objectifs de performance, mais qui pourraient s'avérer utiles à l'avenir. Après qu'un exercice de red teaming a mis en évidence un risque de fuite et de manipulation des données via des détails mémorisés, l'équipe décide de réduire l'exposition. Pour ce faire, elle supprime les champs inutiles et des enregistrements entiers afin que les informations sensibles ne figurent pas dans l'ensemble de données.

Quel contrôle **général** a été mis en œuvre pour limiter les données sensibles dans cette situation ?

A) Données autorisées
B) Minimisation des données
C) Obfuscation des données d'entraînement
D) Conservation de courte durée

A) Incorrect. Les « données autorisées » définissent les catégories autorisées, mais le scénario nécessite la suppression des enregistrements inutiles, ce qui relève de la minimisation des données.
B) Correct. La minimisation des données consiste à supprimer les enregistrements inutiles de l'ensemble d'entraînement afin que les données sensibles ne soient pas présentes et ne puissent donc pas faire l'objet d'une fuite. (Bibliographie : A, chapitre 1)
C) Incorrect. L'obfuscation des données d'entraînement transforme les données mais les conserve. Par conséquent, contrairement à la suppression, des risques peuvent subsister.
D) Incorrect. La conservation de courte durée limite la durée pendant laquelle les données sont conservées, et non les enregistrements inclus dans l'ensemble d'entraînement.

29 / 40
Une entreprise de vente au détail décide de ne plus stocker les informations de paiement de ses clients une fois chaque transaction terminée.

En quoi cette mesure **réduit**-elle le risque d'exposition des données ?

A) En éliminant la nécessité de chiffrer les données en transit ou archivées à des fins d'audit
B) En garantissant la conformité à toutes les réglementations applicables en matière de protection des données, ce qui réduit la complexité
C) En garantissant qu'aucun utilisateur non autorisé ne puisse accéder au réseau et à la base de données de l'entreprise
D) En empêchant que les données conservées ne fassent l'objet de fuites, ne soient reconstituées ou déduites à partir du système

A) Incorrect. La minimisation des données ne rend pas le chiffrement superflu. Les données transmises doivent toujours être protégées, quelle que soit la quantité de données stockées.
B) Incorrect. La minimisation des données soutient les efforts de conformité, mais elle ne garantit pas automatiquement le respect total de toutes les réglementations en matière de protection des données, qui comportent de nombreuses exigences supplémentaires.
C) Incorrect. Limiter les données stockées ne garantit pas que des utilisateurs non autorisés ne puissent pas accéder au réseau ou à la base de données. Le contrôle d'accès au réseau est une mesure de sécurité distincte, sans rapport avec la minimisation des données.
D) Correct. Lorsque les données ne sont ni collectées ni conservées, elles ne peuvent faire l'objet d'une fuite, être reconstituées ou déduites à partir du système. Cela réduit directement l'impact d'un vol de jeux de données ou d'un accès non autorisé. (Bibliographie : A, chapitre 1)

30 / 40
Un réseau hospitalier régional utilise un agent IA pour lire les e-mails entrants des fournisseurs, rédiger des réponses et les envoyer automatiquement. Lors de tests de sécurité menés à la suite d'un récent incident d'hameçonnage, les testeurs ont découvert qu'une injection de prompt pouvait amener l'agent IA à ignorer ses règles et à envoyer des réponses non autorisées.

Quelle décision de conception est la **plus efficace** pour limiter les effets de ce comportement indésirable ?

A) Désactiver toutes les intégrations de l'agent IA et revenir à des workflows entièrement manuels pour la plupart des processus
B) Renforcer la formation des utilisateurs à la sécurité des prompts et mettre l'accent sur la supervision lors de l'envoi d'e-mails assisté par l'agent
C) Conserver des autorisations d'envoi autonomes étendues et passer à un modèle génératif plus avancé
D) Limiter les autorisations de l'agent IA à des périmètres spécifiques à certaines tâches et exiger une validation pour les actions à haut risque

A) Incorrect. Cela supprime l'automatisation au lieu de mettre en place des contrôles permettant de conserver la fonctionnalité en toute sécurité.
B) Incorrect. La sensibilisation des utilisateurs est utile, mais ne permet pas d'appliquer des contrôles en phase d'exécution de l'IA.
C) Incorrect. L'amélioration de la qualité du modèle ne réduit pas suffisamment le rayon d'impact si le système d'IA générative dispose toujours de permissions étendues.
D) Correct. Les contrôles basés sur le moindre privilège et sur l'intervention humaine sont recommandés pour limiter l'impact en cas d'injection de prompt. Le moindre privilège spécifique à chaque tâche limite ce que l'agent peut faire, tandis que l'approbation des actions à haut risque renforce la supervision, empêchant ainsi les réponses non autorisées même lorsque le modèle est trompé. (Bibliographie : A, chapitre 1)

31 / 40
Quelle est la **meilleure** façon de minimiser les effets d'un comportement indésirable du modèle ?

A) Mettre l'accent sur le renforcement des prompts, l'accès basé sur les rôles et les audits programmés des sorties du modèle
B) S'appuyer sur des filtres automatisés des sorties, la journalisation des exceptions et des contrôles ponctuels de l'équité
C) Rechercher les biais indésirables et appliquer des contrôles tels que la supervision, le moindre privilège et la validation continue
D) Recourir à une validation rigoureuse des données d'entrée, à des autorisations d'outils restreintes et à des revues périodiques impliquant une intervention humaine

A) Incorrect. Ces mesures sont utiles mais incomplètes. Le renforcement des prompts est fragile, le contrôle d'accès basé sur les rôles (RBAC) sans programme plus large de moindre privilège est limité, et les audits programmés sont périodiques plutôt que continus, ce qui laisse des lacunes et ne permet pas de tester systématiquement les biais.
B) Incorrect. Les filtres de sortie automatisés peuvent être contournés, la journalisation des exceptions n'empêche pas les préjudices, et les contrôles ponctuels d'équité sont insuffisants par rapport à une validation continue et à des tests structurés de biais.
C) Correct. Les contrôles visant à limiter les effets des comportements indésirables du modèle sont les suivants : supervision, moindre privilège du modèle, alignement du modèle, transparence de l'IA, explicabilité, validation continue et tests de biais indésirables. (Bibliographie : A, chapitre 1)
D) Incorrect. Une validation rigoureuse des données d'entrée, des autorisations restreintes pour les outils et un examen humain périodique sont utiles, mais cette liste omet la validation continue et les tests dédiés aux biais. Les examens périodiques laissent des lacunes et n'assurent pas une réduction complète et continue des risques.

32 / 40
Limiter les comportements indésirables des modèles est une stratégie essentielle pour limiter l'impact et contrôler le rayon d'impact.

Comment cette stratégie **améliore-t-elle** les performances et **réduit-elle** le risque ?

A) En se concentrant sur les attaques susceptibles d'affecter les systèmes d'IA et en limitant la supervision des facteurs de performance et de l'exactitude du système
B) En mettant en œuvre des contrôles qui éliminent la nécessité de tests et de validations continus au fil du temps afin de réduire la complexité
C) En améliorant la fiabilité opérationnelle et l'efficacité des ressources tout en minimisant les incidents et le gaspillage de requêtes de calcul
D) En augmentant l'autonomie du modèle et en lui permettant de générer davantage de sorties afin d'améliorer en continu les performances du système

A) Incorrect. La transparence de l'IA est l'un des contrôles permettant de limiter les comportements indésirables. Elle est directement liée à l'exactitude, car les attaques ou les dérives à l'origine de comportements indésirables constituent des problèmes d'exactitude.
B) Incorrect. Des tests réguliers et une validation des performances garantissent que le modèle reste conforme à l'usage prévu au fil du temps et limitent l'effet des comportements indésirables. Les tests de biais indésirables constituent l'un des contrôles visant à limiter les effets des comportements indésirables dans les systèmes d'IA. Il est fortement recommandé de procéder à des tests et à une validation continus au fil du temps.
C) Correct. La limitation des comportements indésirables réduit les sorties préjudiciables et les incidents opérationnels, améliore la fiabilité et accroît l'efficacité des ressources en réduisant les nouvelles tentatives, les appels d'outils inutiles et le gaspillage de ressources de calcul. (Bibliographie : A, chapitre 1)
D) Incorrect. Accroître l'autonomie du modèle n'entraîne pas d'amélioration des performances. Le niveau d'autonomie du modèle doit être maîtrisé afin de limiter les comportements indésirables du modèle.

33 / 40
Dans le cadre plus large du red teaming (exercice de test adversarial) de l'IA, quel est l'objectif **principal** des tests de sécurité de l'IA ?

A) Évaluer si un modèle d'IA est résilient face à des attaques spécifiques en reproduisant ces attaques dans un environnement contrôlé
B) Évaluer si un modèle d'IA produit des sorties précises et traçables lorsqu'il est exposé à des données d'entrée attendues
C) Mesurer si un modèle d'IA fait preuve de l'efficacité de calcul attendue lors du traitement de grands volumes de requêtes
D) Vérifier si un modèle d'IA répond aux exigences métier définies par le biais de procédures standard d'assurance qualité

A) Correct. Les tests de sécurité de l'IA font partie du red teaming de l'IA et consistent à vérifier si le modèle d'IA peut résister à certaines attaques en simulant ces dernières. Leur objectif principal est donc de vérifier la résilience du modèle par la reproduction contrôlée d'attaques. (Bibliographie : A, chapitre 5)
B) Incorrect. L'évaluation de l'exactitude des sorties dans des conditions d'entrée normales relève des tests fonctionnels ou de performance, et non des tests de sécurité de l'IA, qui se concentrent sur la simulation d'attaques.
C) Incorrect. L'efficacité de calcul dans le cadre de charges de travail courantes relève des tests de performance ou de charge, et non du champ d'application des tests de sécurité de l'IA, axés sur les attaques.
D) Incorrect. La vérification de la conformité aux exigences métier par le biais de l'assurance qualité ne relève pas du champ d'application des tests de sécurité de l'IA, qui ciblent spécifiquement la capacité du modèle d'IA à résister aux attaques.

34 / 40
Une entreprise déploie un assistant d'IA générative et souhaite définir des tests de sécurité de l'IA qui aillent au-delà des simples tests d'intrusion conventionnels. Pour ce faire, l'équipe de test commence par identifier les menaces à tester.

Quelles menaces **spécifiquement** associées aux systèmes d'IA générative devraient être incluses dans les tests de sécurité de l'IA ?

A) Attaques par évasion, gestion non sécurisée des sorties et empoisonnement du modèle
B) Exfiltration du modèle, injection de prompt et attaques par évasion
C) Injection de prompt, divulgation de données sensibles et gestion non sécurisée des sorties
D) Sortie de données sensibles par le modèle, dérive du modèle et invocation non sécurisée d'outils

A) Incorrect. Les attaques d'évasion sont plus couramment associées aux systèmes d'IA prédictive, tandis que la gestion non sécurisée des sorties peut affecter les systèmes d'IA générative. Cette combinaison ne représente pas au mieux les principales menaces spécifiques à l'IA générative qui devraient être incluses dans les tests de sécurité de l'IA.
B) Incorrect. Bien qu'il s'agisse de menaces de sécurité valides pour l'IA, les attaques d'évasion sont plus couramment associées aux systèmes d'IA prédictive qu'aux assistants d'IA générative.
C) Correct. Ces trois menaces affectent l'IA générative et doivent donc faire l'objet de tests. (Bibliographie : A, chapitre 5)
D) Incorrect. La dérive du modèle est un problème de performance, et non une menace de sécurité.

35 / 40
Une équipe de sécurité de l'IA teste la résistance à l'injection de prompt d'un agent conversationnel (chatbot) d'IA générative. Au cours du test, le système a bloqué plusieurs entrées d'attaque spécialement conçues.

Quelle est la **prochaine** étape recommandée pour continuer à assurer la supervision de la sécurité de l'agent conversationnel d'IA générative ?

A) Ajouter des variations dans les entrées, telles que le remplacement par des synonymes, des modifications d'encodage ou des changements de mise en forme, afin de tenter de contourner la détection
B) Conclure le cycle de test en cours et considérer que les défenses sont adéquates, puisque la plupart des tentatives d'injection de prompt ont été bloquées
C) Répéter le même test plusieurs fois afin d'augmenter la taille de l'échantillon et d'obtenir des données plus pertinentes sur le plan statistique concernant les performances des défenses
D) Envoyer les entrées d'attaque par injection de prompt directement au modèle afin de mieux l'entraîner contre des menaces similaires lors de futurs déploiements

A) Correct. L'ajout d'algorithmes de variation, tels que les synonymes, l'encodage et la mise en forme, pour tenter de contourner la détection renforce la robustesse de celle-ci. Élargir la diversité des entrées améliore la couverture, met en évidence les failles de contournement et maintient la supervision de la sécurité. (Bibliographie : A, chapitre 5)
B) Incorrect. Le test ne doit pas encore être considéré comme terminé. Le système doit encore faire preuve de résilience face à des tentatives délibérées de contournement.
C) Incorrect. Répéter les mêmes tests sans introduire de variation a peu de chances de révéler de nouvelles techniques de contournement ou d'améliorer la couverture.
D) Incorrect. L'envoi d'entrées d'attaque brutes directement au modèle contourne les filtres de production et les mécanismes de détection. De plus, l'intégration directe d'entrées d'attaque dans l'entraînement relève de l'amélioration du modèle, et non de l'étape suivante des tests.

36 / 40
Le système d'IA d'une entreprise s'appuie sur de grandes quantités de données à caractère personnel pour l'entraînement et la prise de décision.

D'un point de vue éthique, quel est le risque **le plus** important que l'entreprise doive traiter ?

A) La collecte, la conservation ou la réutilisation des données au-delà des finalités justifiées
B) La complexité de l'interface rendant difficile la vérification des décisions automatisées
C) La baisse de l'exactitude du modèle lorsque les ressources sont réparties de manière inégale
D) Les mises à jour logicielles entraînant des problèmes de compatibilité entre les plateformes

A) Correct. Les systèmes d'IA sont gourmands en données et soulèvent des inquiétudes quant à la manière dont les données à caractère personnel sont collectées, conservées et réutilisées. Les principes et la législation en matière de protection de la vie privée exigent souvent une base légale ou un consentement pour une telle utilisation, ainsi que des garanties telles que la limitation de l'utilisation et la minimisation des données. (Bibliographie : A, chapitre 6)
B) Incorrect. La complexité de l'interface relève principalement de l'ergonomie. Elle ne traite pas directement des risques liés aux systèmes d'IA s'appuyant sur de grandes quantités de données à caractère personnel.
C) Incorrect. L'allocation des ressources matérielles concerne les performances techniques et le fonctionnement des modèles d'IA. Ce n'est pas la principale préoccupation liée à l'utilisation des données à caractère personnel dans les systèmes d'IA.
D) Incorrect. Les problèmes de mise à jour et de compatibilité des logiciels concernent la maintenance et l'interopérabilité du système. Ils ne constituent pas la principale préoccupation liée aux systèmes d'IA utilisant de grandes quantités de données à caractère personnel.

37 / 40
Une librairie en ligne utilise un système d'IA pour analyser l'historique de navigation, les tendances d'achat et les précommandes afin de prévoir la demande et de s'assurer que les livres populaires soient disponibles à temps. Par la suite, l'entreprise réutilise ces mêmes données clients pour entraîner un autre modèle d'IA qui établit des profils marketing détaillés et envoie des publicités ciblées à chaque acheteur.

Quel principe de protection de la vie privée est **le plus directement** enfreint ici ?

A) Minimisation des données
B) Limitation de la conservation des données
C) Transparence et consentement
D) Limitation de l'utilisation et spécification de la finalité

A) Incorrect. La minimisation des données consiste à ne collecter que les données strictement nécessaires à une finalité définie, et non à restreindre la réutilisation de données déjà collectées à d'autres fins.
B) Incorrect. La limitation de la durée de conservation des données concerne la durée pendant laquelle les données sont conservées avant de devoir être supprimées, et non la question de savoir si des données collectées dans un but donné peuvent être réutilisées à d'autres fins.
C) Incorrect. La transparence et le consentement concernent l'information des personnes sur la collecte de données et l'obtention de leur consentement, et non spécifiquement la restriction relative à l'utilisation de données collectées dans un but donné pour entraîner un modèle à des fins différentes.
D) Correct. Ce scénario illustre une violation des principes de limitation de l'utilisation et de spécification de la finalité, car les données collectées à des fins de sûreté et de sécurité ne devraient pas être réutilisées comme ensemble de données d'entraînement pour le profilage ou le marketing personnalisé. (Bibliographie : A, chapitre 6)

38 / 40
Une organisation souhaite mettre en place un cadre pour régir l'utilisation responsable de l'IA dans l'ensemble de ses activités.

Elle utilise la norme ISO/IEC 27001 pour le management de la sécurité de l'information. L'équipe chargée de la sécurité souhaite mettre en place sa gouvernance de l'IA à l'aide d'une norme ISO/IEC qui remplisse, pour la gouvernance de l'IA, le même rôle structurel que la norme ISO/IEC 27001 pour la sécurité de l'information.

De quelle norme s'agit-il ?

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

A) Incorrect. La norme ISO/IEC 23894 fournit des recommandations en matière de gestion des risques liés à l'IA, et non un système de management destiné à la gouvernance d'une IA responsable.
B) Incorrect. La norme ISO/IEC 27005 traite de la gestion des risques liés à la sécurité de l'information et n'est pas conçue pour servir de système de management de la gouvernance d'une IA responsable.
C) Correct. La norme ISO/IEC 42001 peut être considérée comme une norme de système de management pour la gouvernance d'une IA responsable au sein d'une organisation, à l'instar de la norme ISO/IEC 27001 qui est une norme de système de management pour la sécurité de l'information. (Bibliographie : A, chapitre 1)
D) Incorrect. La norme ISO/IEC 5338 vise à étendre les pratiques et l'ingénierie du cycle de vie des logiciels, et non à fournir un système de management de la gouvernance pour une IA responsable.

39 / 40
Une entreprise souhaite utiliser un système d'IA pour sélectionner et classer automatiquement les candidats à un poste, sur la base de leurs CV.

Selon le règlement sur l'IA, dans quelle catégorie l'utilisation de ce système d'IA doit-elle être classée ?

A) Risque inacceptable
B) Haut risque
C) Risque limité
D) Risque minimal ou nul

A) Incorrect. Le règlement sur l'IA n'interdit pas les outils de recrutement basés sur l'IA. Il les classe dans la catégorie « haut risque », ce qui signifie qu'ils sont soumis à des exigences et restrictions spécifiques, mais qu'ils ne sont pas purement et simplement interdits.
B) Correct. La sélection des candidats à l'embauche peut avoir une incidence sur l'accès à l'emploi et les droits fondamentaux et doit être classée dans la catégorie « haut risque ». Les organisations doivent s'assurer de la conformité avant de déployer des systèmes d'IA à ces fins. (Bibliographie : A, chapitre 6)
C) Incorrect. Le risque limité concerne principalement les obligations liées à la seule transparence. La sélection à l'embauche peut avoir une incidence sur l'accès à l'emploi et les droits fondamentaux et doit être classée dans la catégorie « haut risque ».
D) Incorrect. Le risque minimal concerne les utilisations sans conséquence, comme les filtres anti-spam ou les jeux vidéo. La sélection à l'embauche peut avoir une incidence sur l'accès à l'emploi et les droits fondamentaux et doit être classée dans la catégorie « haut risque ».

40 / 40
Une équipe de développement en IA prépare des données d'entraînement pour un nouveau modèle d'apprentissage automatique. Le chef d'équipe souhaite s'assurer que l'approche de collecte des données favorise la conformité en matière de droits d'auteur.

Quelle est la pratique **la plus éthique** en matière de collecte de données ?

A) Collecter des données uniquement à partir de sources accessibles au public
B) Créer l'ensemble des données d'entraînement au sein de l'organisation
C) Télécharger des données provenant de sources dont l'autorisation est implicite
D) N'utiliser que des données largement partagées en ligne

A) Incorrect. Le fait que les données soient accessibles au public ne garantit pas à lui seul qu'elles soient sous licence pour l'usage prévu par l'organisation. Les conditions de licence doivent être examinées afin de confirmer qu'elles sont suffisantes pour l'application prévue.
B) Correct. La production de données en interne est l'une des stratégies éthiques reconnues en matière de collecte de données, car elle garantit à l'organisation la pleine maîtrise des données utilisées pour entraîner les modèles d'IA. (Bibliographie : A, chapitre 0)
C) Incorrect. L'autorisation implicite ne constitue pas un fondement valable pour l'utilisation des données. Des autorisations explicites doivent être obtenues auprès du propriétaire des données avant d'utiliser des données de tiers pour l'entraînement de l'IA.
D) Incorrect. Un contenu largement republié en ligne peut tout de même être protégé par le droit d'auteur. La répétition ou la large diffusion ne rend pas le contenu libre de réutilisation pour l'entraînement de l'IA. Il convient de se concentrer sur le fait de savoir si les données sont créées en interne, correctement concédées sous licence ou obtenues avec les autorisations nécessaires.

Évaluation

Le tableau ci-dessous indique les bonnes réponses aux questions de cet exemple d'examen.

Question	Réponse	Question	Réponse
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

Contacter EXIN

www.exin.com