|  |  |  |  |
| --- | --- | --- | --- |
| Réflexion 2 – Identifier les risques soulevés par les IA | | | |
| Durée : 20’ | Homme avec un remplissage uniouDeux hommes avec un remplissage uni | Une image contenant texte, symbole, Police, Graphique  Le contenu généré par l’IA peut être incorrect. | Source |

**Travail à faire**

Après avoir lu le document, répondez aux questions suivantes

1. Quelles sont les quatre grandes catégories de risques liés à l’utilisation de l’IA identifiées dans le texte ?
2. Donnez un exemple de risque de sécurité ou de confidentialité présenté dans le document.
3. En quoi les « hallucinations » générées peuvent-elles être problématiques dans certains secteurs ?
4. Pourquoi le développement interne de modèles de GenAI représente-t-il un risque opérationnel et financier pour les entreprises ?

**Doc.  10 types de risques liés à l'IA générative et stratégies d'atténuation**

*Source : https://layerxsecurity.com/ - 11 juin 2024 (synthèse de l’article)*

L’utilisation généralisée de l’**IA générative** (*GenAI*) dans tous les secteurs exige une conscience des **risques sécuritaires et opérationnels** ainsi que des options d'atténuation. Cet article présente les **10 principaux risques**.

L’utilisation d’applications GenAI et de LLM (publiques ou internes) présente des risques majeurs pour les organisations, répartis en quatre catégories :

**Catégorie n°1 : Risques de sécurité et de confidentialité**

1. **Problèmes de confidentialité** : La GenAI s'appuie sur de vastes données d'entraînement, incluant potentiellement des informations personnelles sensibles. L'utilisation de ces données dans les résultats peut entraîner des violations de la vie privée. La nature « boîte noire » de nombreux modèles complique la transparence.
2. **E-mails de phishing et logiciels malveillants** : La GenAI permet aux cybercriminels de créer des attaques de **phishing** très convaincantes, imitant le style de communications légitimes, et de développer des **logiciels malveillants** évolutifs contournant les sécurités traditionnelles.
3. **Menaces internes et abus des employés** : Les employés (initiés) peuvent exploiter leur accès pour un vol ou un sabotage intentionnel, ou par négligence, en **saisissant par inadvertance des données sensibles** (code source, informations client, etc.) dans les applications GenAI.
4. **Surface d'attaque accrue** : Les systèmes GenAI augmentent la surface d'attaque en s'intégrant à diverses sources de données et API, créant de multiples points d'entrée que des acteurs malveillants peuvent exploiter (ex : injection de données malveillantes).

**Catégorie n°2 : Risques liés à la qualité et à la fiabilité**

1. **Problèmes de qualité de sortie** : L'IA peut générer des résultats **inexacts, trompeurs ou biaisés**, dus à des données de formation inadéquates ou à l'imprévisibilité des algorithmes. Dans les applications critiques (santé, finance), cela peut entraîner de graves pertes ou mettre des vies en danger.
2. **« Faits » et hallucinations inventés** : Un risque extrême est l'**hallucination**, où le LLM génère des informations **fausses mais plausibles**, basées sur des modèles de données plutôt que sur l'exactitude factuelle. Cela risque de propager des informations incorrectes, notamment dans des secteurs où l'exactitude est vitale (droit, finance).

**Catégorie n°3 : Risques juridiques et éthiques**

1. **Droits d'auteur, propriété intellectuelle et autres risques juridiques** : L'entraînement des modèles sur de grandes quantités de données, souvent protégées par le **droit d'auteur**, peut entraîner une reproduction non intentionnelle de contenu protégé. La légalité de cet entraînement et l'originalité des contenus générés sont actuellement débattues et font l'objet de litiges juridiques (ex : poursuites contre OpenAI/Microsoft).
2. **Résultats biaisés** : Les systèmes GenAI peuvent produire des **résultats biaisés** si les données d'entraînement sont non représentatives, reflétant des inégalités. Cela peut mener à des pratiques **discriminatoires** (embauche, prêts, santé), perpétuant les préjugés sociétaux.
3. **Conformité** : Le traitement d'informations sensibles par des systèmes GenAI expose au risque de **fuite de données** et de non-conformité aux réglementations, surtout si le fournisseur n'a pas de mesures de sécurité robustes.

**Catégorie n°4 : Risques opérationnels et financiers**

1. **Coût de l'expertise et du calcul** : Le développement et le déploiement internes de LLM exigent des **dépenses considérables** pour les GPU hautes performances, le matériel spécialisé et les services *cloud*. De plus, les professionnels qualifiés (data scientists, ingénieurs ML) perçoivent des **salaires élevés**. La pénurie mondiale accentue ces coûts, créant une barrière à l'entrée.

**Réponses**

1. **Quelles sont les quatre grandes catégories de risques liés à l’utilisation de l’IA identifiées dans le texte ?**
2. **Donnez un exemple de risque de sécurité ou de confidentialité présenté dans le document.**
3. **En quoi les « hallucinations » générées peuvent-elles être problématiques dans certains secteurs ?**
4. **Pourquoi le développement interne de modèles de GenAI représente-t-il un risque opérationnel et financier** pour les entreprises ?