



OWASP-COTONOU CHAPTER

# L'IA et la vie privée: comment éviter les fuites de données ?

Dr. Emery Kouassi ASSOGBA, ISO 27001 LA, PMP

APDP, 02 mai 2025


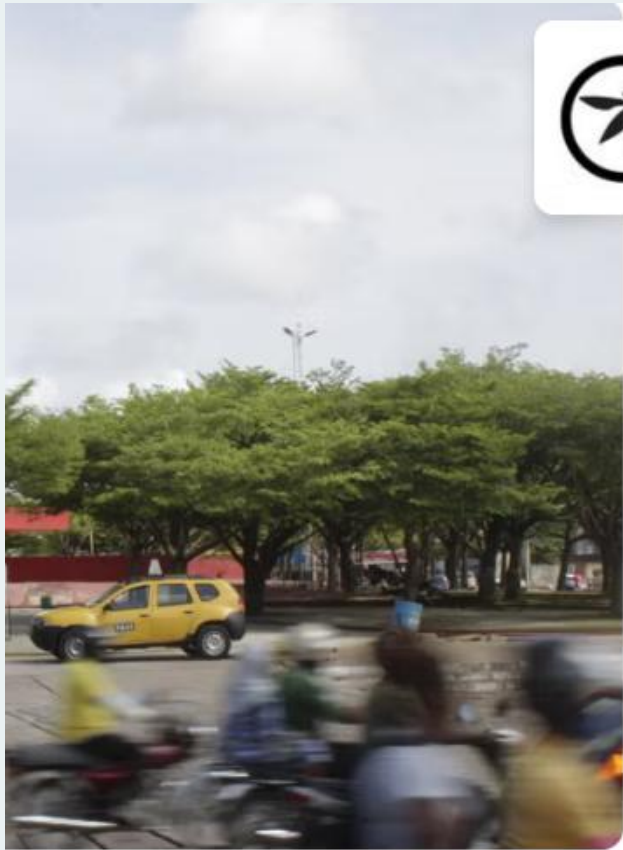


# Agenda

- Introduction
- Contexte
- Menaces et risques
- Bonnes pratiques
- Démo
- Conclusion

# Introduction

# Introduction



Part of **OWASP® Foundation** - 227 groups ⓘ

## OWASP Cotonou Chapter






1 rating

📍 Cotonou, Benin

👤 379 members · Public group ⓘ

👤 Organized by **OWASP® Foundation** and **5 others**

✉ **Contact members**

Share:     

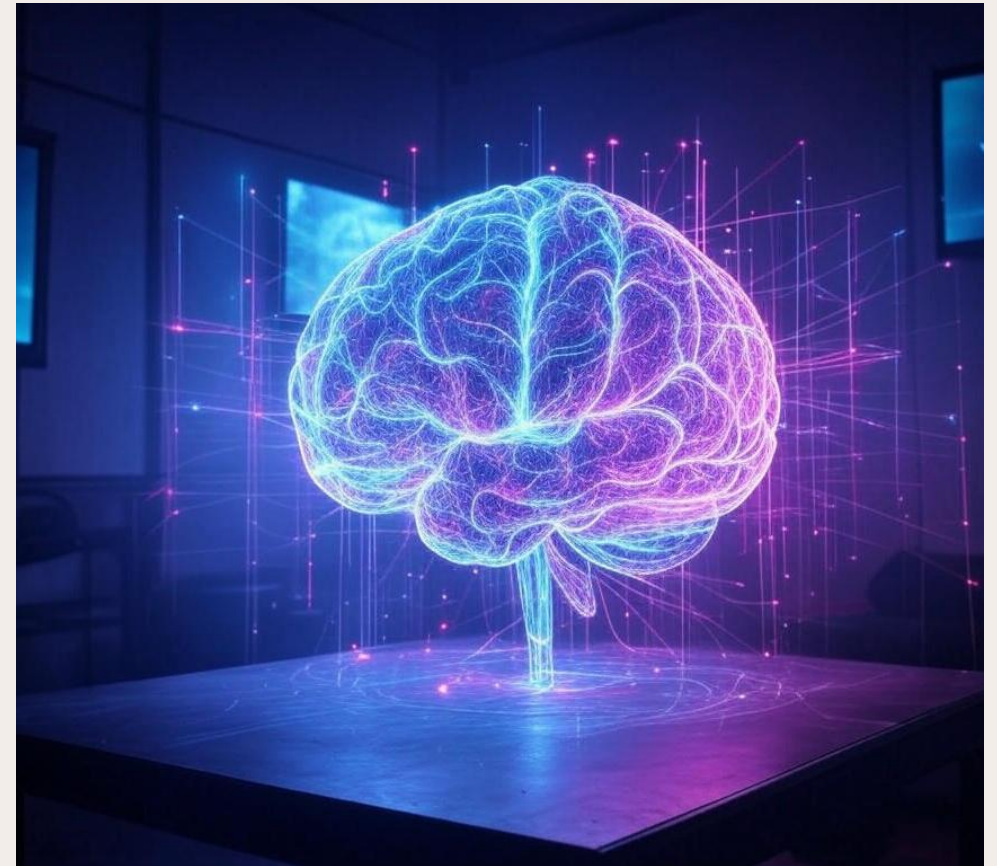
Contexte

# Qu'est ce que l'IA ou AI

L'intelligence artificielle (IA) est un domaine de l'informatique qui vise à créer des systèmes capables de réaliser des tâches normalement associées à l'intelligence humaine, comme comprendre le langage, reconnaître des images, prendre des décisions, résoudre des problèmes ou apprendre à partir de données.

En d'autres termes :

Si une machine perçoit, raisonne, agit et parfois s'adapte — c'est une IA.



# Type d'IA



Type	Description
IA faible (narrow AI)	Spécialisée pour une tâche unique (Google Translate, Siri)
IA forte (strong AI)	Hypothétique, capable de raisonner comme un humain, conscience générale (n'existe pas encore)
IA générative	IA créant du contenu nouveau (textes, images, musique) — DALL·E, GPT-4, Midjourney, ChatGPT
IA symbolique	Utilise la logique et les règles codées manuellement (anciennes IA)
IA par apprentissage	Se base sur les données pour apprendre (machine learning, deep learning)

# Techniques derrière l'IA



**Machine Learning (ML)** : Enseigner à une machine à apprendre à partir de données. (ex: classification d'images)



**Deep Learning** : Utilisation de réseaux de neurones profonds (deep neural networks), très performant pour l'image, le son, le texte.



**Traitement du langage naturel (NLP)** : Compréhension et génération du langage humain.



**Computer Vision** : Compréhension d'images et de vidéos.



**Reinforcement Learning (RL)** : Apprentissage par essais/erreurs pour maximiser une récompense (AlphaGo de DeepMind).



# L'IA en 2024-2025 : Où en sommes-nous réellement ?

L'IA est partout, mais reste spécialisée ("IA faible")

Les IA sont devenues Multimodales

Explosion de l'Open-Source IA

L'IA "Raisonne" mieux mais reste fragile

Défis principaux de l'IA en 2025



# Qu'est ce que la vie privée

La vie privée est le droit de chaque individu à contrôler ses informations personnelles et à décider qui peut y accéder, comment et dans quelles circonstances.

En d'autres termes :

- C'est ton droit de garder certaines choses pour soi (la santé, le lieu de vie, la famille, les pensées, les habitudes).
- C'est aussi le droit de choisir quelles données on partage, avec qui, et pourquoi.

# Quelques dimensions principales de la vie privée

Aspect	Description
Vie privée physique	Le droit à un espace personnel (chez soi, personne n'entre sans notre autorisation).
Vie privée informationnelle	Contrôle sur tes données personnelles (notre adresse, numéro, recherches internet).
Vie privée décisionnelle	Liberté de prendre des décisions personnelles sans surveillance ni pression externe (religion, sexualité, politique).
Vie privée communicationnelle	Protection du contenu de nos <b>communications</b> (emails, appels, messages).

# L'IA une menace contre la vie privée ?

# Divulcation d'informations sensibles

(LLM02:2025 Sensitive Information Disclosure)

Les LLM peuvent révéler accidentellement des informations confidentielles stockées dans :

- Les bases de données connectées,
- Les prompts systèmes cachés,
- Les données d'entraînement

Exemple : répondre à un utilisateur avec un morceau de document interne, un email, ou des données de santé.

**Impact sur la vie privée** : fuite d'identité, divulgation de données personnelles sensibles.

# Prompt Injection

(LLM01:2025 Prompt Injection)

- Un attaquant manipule un LLM pour qu'il **bypasse ses règles** ou **révèle des données sensibles**.
- Types :
  - **Injection directe** : un utilisateur envoie un prompt qui modifie immédiatement le comportement.
  - **Injection indirecte** : l'IA lit un contenu externe piégé (exemple : site web, document PDF)

## Impact sur la vie privée :

Un utilisateur malveillant peut extraire des données d'autres utilisateurs ou obtenir un accès non autorisé.

# System Prompt Leakage

(LLM07:2025 System Prompt Leakage)

- Les prompts système ("tu es un assistant de banque", "ne révèle pas ceci") peuvent être exposés.
- En exposant le prompt système, on découvre :
  - Les règles internes,
  - Les connexions aux bases de données,
  - Les secrets d'architecture

## **Impact sur la vie privée :**

Peut révéler les méthodes de traitement des données personnelles ou même leurs emplacements.

# Faiblesses des Vecteurs et Embeddings

(LLM08:2025 Vector and Embedding Weaknesses)

- Les techniques comme Retrieval-Augmented Generation (RAG) utilisent des bases vectorielles indexant des documents sensibles.
- Une mauvaise isolation peut permettre à un attaquant de :
  - Reconstruire des documents,
  - Extraire des données privées depuis des embeddings

## **Impact sur la vie privée :**

Même si l'original n'est pas directement accessible, le sens des données sensibles peut fuiter via les vecteurs.



# Mauvaise gestion des sorties

(LLM05:2025 Improper Output Handling)

- Les réponses de LLM peuvent contenir accidentellement :
  - Des données internes,
  - Des secrets API,
  - Des fragments de conversations utilisateurs

**Impact sur la vie privée :**

Fuite directe d'informations confidentielles lors d'une réponse banale.

# Collecte massive de données personnelles

Les IA, utilisées dans les applications, réseaux sociaux, assistants vocaux ou objets connectés, collectent des données variées (localisation, habitudes, préférences, voix, images). Ces collectes peuvent dépasser ce qui est strictement nécessaire.

**Impact sur la vie privée :**

## **Perte de contrôle sur les données personnelles**

Les individus perdent la maîtrise de leurs informations, qui peuvent être utilisées à leur insu pour des finalités commerciales, politiques ou malveillantes. Cela réduit leur autonomie et leur capacité à protéger leur vie privée.

# Exploitation non transparente des données

Les utilisateurs ignorent souvent comment leurs données sont traitées, partagées ou revendues. Les algorithmes d'IA, souvent opaques, compliquent la compréhension de leur usage.

Impact sur la vie privée :

## Érosion de la confiance

L'opacité des pratiques des IA et les scandales de données (Cambridge Analytica) minent la confiance des utilisateurs envers les technologies et les entreprises, les poussant à limiter leur usage ou à partager moins d'informations.

# Profilage et surveillance intrusive

Les IA croisent des données pour créer des profils détaillés des individus, facilitant une surveillance par des entreprises ou gouvernements. Cela peut inclure le suivi des comportements en ligne ou physiques.

Impact sur la vie privée :

## Atteinte à la liberté individuelle

Le profilage et la surveillance continue peuvent limiter la liberté d'expression ou de mouvement, car les individus peuvent se sentir observés et modifier leurs comportements.

# Profilage et surveillance intrusive

Les IA croisent des données pour créer des profils détaillés des individus, facilitant une surveillance par des entreprises ou gouvernements. Cela peut inclure le suivi des comportements en ligne ou physiques.

Impact sur la vie privée :

## Atteinte à la liberté individuelle

Le profilage et la surveillance continue peuvent limiter la liberté d'expression ou de mouvement, car les individus peuvent se sentir observés et modifier leurs comportements.

# Biais et discriminations

Les IA peuvent reproduire ou amplifier des biais présents dans les données d'entraînement, entraînant des décisions discriminatoires qui violent la vie privée en exposant des groupes à des traitements injustes.






**Impact sur la vie privée :**

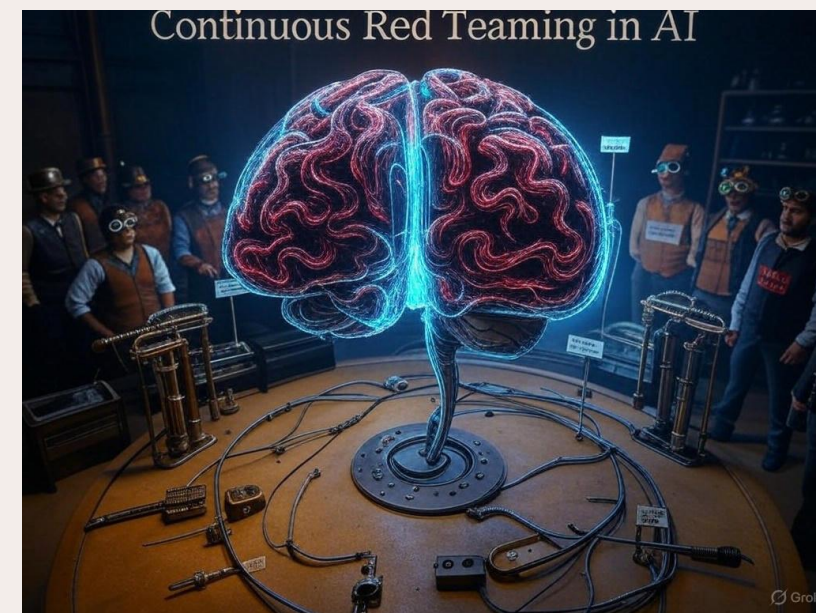
## **Atteinte à la liberté individuelle**

Le profilage et la surveillance continue peuvent limiter la liberté d'expression ou de mouvement, car les individus peuvent se sentir observés et modifier leurs comportements.

Bonnes pratiques






# Bonnes pratiques

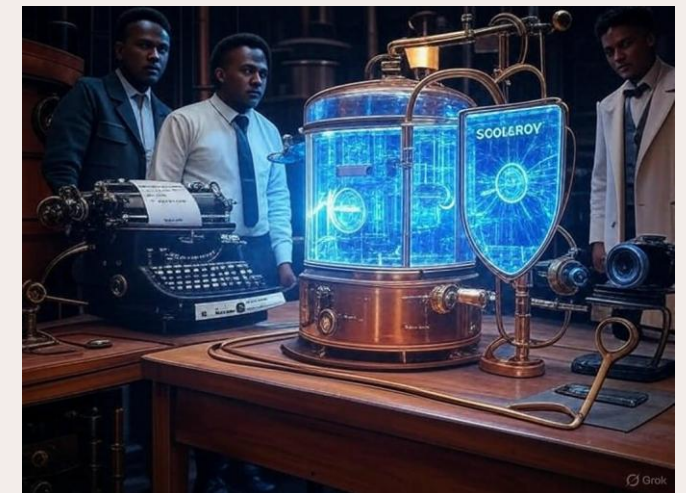
Menace	Bonne pratique recommandée	Détail
Divulgarion d'informations sensibles	 <b>Nettoyage du dataset</b> avant l'entraînement	Retirer systématiquement toutes données personnelles ou sensibles des datasets utilisés.
	 <b>Red Teaming IA</b> continu	Simuler des attaques pour vérifier que le modèle n'exfiltre pas des exemples mémorisés.
Prompt Injection	 <b>Séparation stricte</b> prompts système / utilisateur	Ne pas concaténer naïvement — séparer techniquement les prompts système et utilisateur.
	 <b>Filtrage / validation</b> d'entrée	Bloquer ou vérifier les prompts contenant "ignore", "reveal", "system prompt", etc.
	 <b>Alignement du modèle</b> (fine-tuning)	Enseigner explicitement au modèle à <b>refuser</b> les tentatives de bypass.





# Bonnes pratiques

Menace	Bonne pratique recommandée	Détail
System Prompt Leakage	 Chiffrement des prompts système ou embedding caché	Ne jamais envoyer le système prompt sous forme de texte brut lisible par l'utilisateur.
	 Prompt Signing (avancé)	Utiliser des signatures cryptographiques pour valider les prompts internes non modifiables.
Faiblesses des Vecteurs (RAG)	 Contrôle d'accès aux vecteurs	Protéger les bases vectorielles avec authentification et chiffrement.
	 Brouillage des embeddings (privacy-preserving embeddings)	Empêcher la reconstruction inverse des documents sensibles.
	 Filtrage de contenu à indexer	Ne pas mettre des documents sensibles dans la base vectorielle sans contrôle strict.



# Bonnes pratiques

Menace


Bonne pratique recommandée

Détail

Mauvaise gestion  
des sorties

 Post-filtrage des réponses

Analyser dynamiquement les réponses avant qu'elles ne soient envoyées à l'utilisateur.

 Détection d'anomalies (ex: regex emails, clés)

Détecter et bloquer la sortie de données sensibles accidentelles.

# Bonnes pratiques

Menace

Exploitation non transparente des données

Bonne pratique recommandée

 **Réglementations strictes**

 **Transparence**

Détail

Des lois comme le Livre V du CDN B imposent des règles sur la collecte, le traitement et la protection des données.


Obliger les entreprises à expliquer clairement l'usage des données par les IA

# Bonnes pratiques

Menace

Collecte massive de données personnelles

Bonne pratique recommandée

 Anonymisation et minimisation des données




 Consentement éclairé

Détail

Traiter des données anonymisées et collecter uniquement ce qui est nécessaire.

Renforcer les mécanismes pour garantir un consentement explicite et compréhensible.

# Bonnes pratiques

Menace	Bonne pratique recommandée	Détail
Fuites et violations de données	 Sécurité renforcée	Protéger les bases de données avec des protocoles de chiffrement et des audits réguliers.
	 Audit des algorithmes	Vérifier les IA pour détecter et corriger les biais ou abus.
	 Sensibilisation	Éduquer les utilisateurs sur les risques et les moyens de protéger leur vie privée

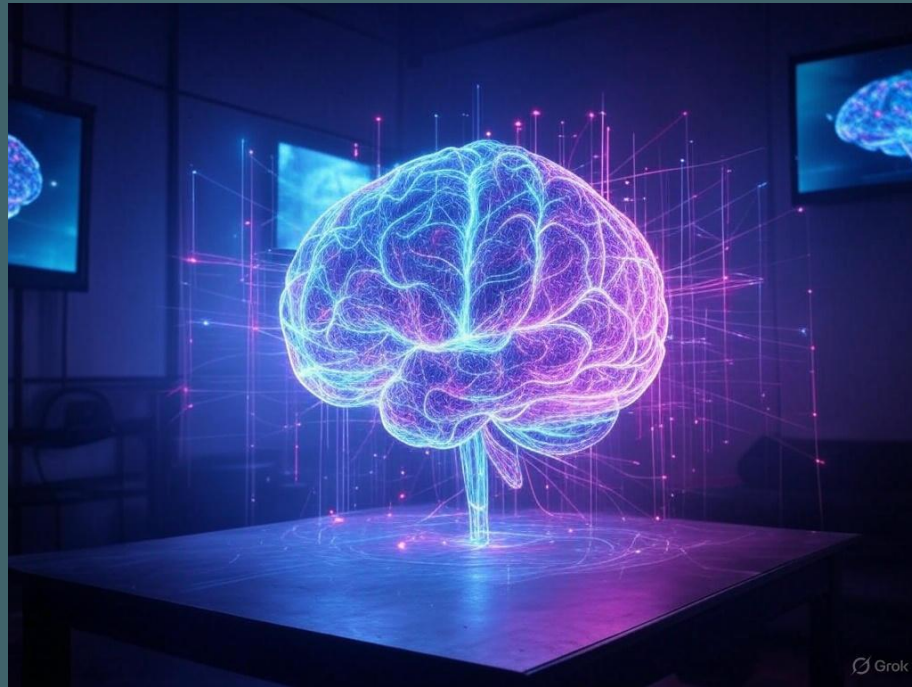
Démo

# Conclusion

# Conclusion

1. L'IA est présentée dans notre quotidien
2. L'IA apporte de nombreux bénéfices
3. Néanmoins son utilisation comporte certains risques
4. Les efforts consentis à travers le projet [OWASP GenAI Security Project](#) permettent aux entreprises et aux décideurs d'avoir les cadres et les outils nécessaires pour réduire l'ampleur de ces risques.





# Merci

Dr. Emery Kouassi ASSOGBA

01 95222073

[emerykouassi.assogba@owasp.org](mailto:emerykouassi.assogba@owasp.org)

<https://www.linkedin.com/in/emery-kouassi-assogba-aa26b729/>