FORMATION

OWASP®
CHAPTER COTONOU

MODULE 4:

PERSPECTIVES FUTURES ET ÉTHIQUE DE L'IA GENERATIVE

N. Janvier AHOUANSOU,
Ing. Sécurité Informatique
Auditeur qualifié GPSE

PARTENAIRES

EMIS sarl

APDP
AUTORITÉ DE PROTECTION DES DONNÉES PERSONNELLES

ASIN
AGENCE DES SYSTÈMES D'INFORMATION ET DU NUMÉRIQUE
RÉPUBLIQUE DU BÉNIN

INSTINCT LEADER

PLAN

Introduction

1. Tendances 2026 : AI vs. AI dans les guerres cyber
2. Éthique OWASP : Biais dans les modèles IA et régulations
3. Roadmap personnel : Intégrer l'IA dans les audits sécurité

Questions / Réponses

Conclusion

... INTRODUCTION

L’IA est devenue une force sismique transformant tous les secteurs essentiels de la vie: industrie, communication, recherche, santé, finance, sécurité, ...

Malgré l’opportunité relative qu’apporte l’IA, des préoccupations persistes: le **biais algorithmique**, la **désinformation à grande échelle**, la **perte de contrôle sur les décisions; la maîtrise de la qualité des données, les données synthétiques**, et la **redéfinition même de ce qu'est le travail humain**.

Quel sera le futur de ce nouveau monde sans repère éthique ni socle réglementaire solide, où la capacité d’un modèle peut être utiliser à la fois pour attaquer et pour se défendre ?

La présentation de ce jour, à l’occasion de la formation OWASP traitera des « Perspectives futures: éthique et Régulation de l’IA ».

Nous parlerons :

- Des tendances 2026 de l’IA;
- **L'Éthique :** Qui doit définir les principes moraux de l'IA ? Comment intégrer la *valeur humaine* au cœur du code ?
- **La Régulation :** Quelle est la place des législateurs ? (comme l'**AI Act** en Europe, et les initiatives dans le reste du monde). »
- De roadmap personnel « feuille de route personnelle » pour les acteurs des fonctions IT

1. Tendances 2026 : AI vs AI dans les guerres cyber

Les mois à venir verront émerger plusieurs tendances dans l'usage et le recours à l'IA soit sous une forme défensive ou offensive.

Déjà, les présentations ont montré que les cyberattaques ne sont plus seulement automatisées : elles deviennent **cognitives**, orchestrées par des IA capables de planifier, d'apprendre et de s'adapter à des changements.

En réponse aux usages offensifs, les défenses utilisent aussi des IA au plan éthique pour contrer les agissements des, créant un **champ de bataille numérique où les deux camps apprennent l'un de l'autre en temps réel**

❑ : Adaptation des attaques pilotées par IA

- Attaques testant automatiquement plusieurs stratégies jusqu'à trouver une vulnérabilité à exploiter.
- Systèmes capables d'analyser les patchs et les signatures des mesures de sécurité déployées.
- IA multimodales utilisée pour exploiter les erreurs humaines (hameçonnage vocal, visuel, textuel).

Les IA offensives génèrent **des attaques entièrement sur mesure**, basées sur une modélisation comportementale du système ciblé.

❑ Autonomisation de la cyber défensive (AI Blue Team)

- Systèmes de détection avec **analyse comportementale continue**.
- IA capables de **neutraliser automatiquement** certains vecteurs d'attaque.

Réponse dynamique : isoler des machines, reconfigurer des réseaux, limiter des privilèges.

Les SOC devront adopter des agents IA pour :

- Une surveillance continue;
- Une corrélation des milliers d'indicateurs
- réagissent en temps réel
- Mettre en place une politique dynamique en fonction des nouvelles données (CVE, …)

❑ Désinformation automatisée et contreattaques IA

- Deepfakes multi-modaux (voix + vidéo + documents) qui va se performer;
- Génération massive de faux contenus ciblés
- Détection automatisée par IA spécialisée (analyse de source, watermarking, cohérence sémantique).

❑ Montée des “agents cyber autonomes

- scanning autonome de vulnérabilités,
- cartographie automatique d’infrastructures,
- coordination entre agents;
- simulation de scénarios cyber anticipatifs.

La plateforme **NodeZero** (de Horizon3.ai) utilise déjà l'IA pour aller au-delà du simple scanning.
Après l’identification de vulnérabilité, l'agent IA **détermine de manière autonome comment l'exploiter** et planifie une **chaîne d'attaque** pour évaluer le véritable impact sur le réseau.
La criticité de l’impact n’est plus seulement appréciée par rapport au CVSS mais réellement

❑Convergence IA-IoT-cyber pour faire des attaques physiques

L'IA rend possibles :

. attaques sur infrastructures critiques,
. manipulations de capteurs ou de flux industriels,
. intrusions dans systèmes autonomes (drones, robots).

Les défenses intègrent donc des modèles IA spécialisés dans :

. la détection d'anomalies physiques,
. la protection des infrastructures OT (Operational Technology).

Ver Stuxnet en 2010, sur les installations nucléaires iraniennes, qui depuis le réseau l'attaque classique se propage à l'API de l'automate industriel de traitement de l'eau créant sa destruction physique

❑Course mondiale à l'IA défensive souveraine

Au niveau État, ils renforcent leurs :

. capacités de cyberdéfense nationale (CERT, CSIRT, SOC,),
. cloud souverains sécurisés,
. IA de détection stratégique,
. unités militaires hybrides : cyber + IA.

Au plan régional, des organisations citoyennes travaillent déjà à la souveraineté des données traitées via les cloud et la collecte excessive des données par les outils d'IA

2. Éthique OWASP : Biais dans les modèles IA et régulations

Les enjeux éthique et réglementaire de l'AI cyber soulèvent des problématiques diversifiées.

Au plan éthique, il se pose souvent la question de:

- Existe-t-il un organe de contrôle l'autonomie des IA offensives / défensives ?
- Comment éviter l'emballement algorithmiques dans les conflits ?
- Quelle responsabilité en cas d'escalade provoquée par une IA mal alignée ?

Sur le plan légal, réglementaire et normatif:

- Quelles normes internationales traitent de l'autonomie cyber,
- La transparence des systèmes défensifs, le principe de l'auditabilité, de traçabilité,
- L'obligation de faire superviser par l'homme ("Human in the Loop").

2. 1. Éthique OWASP : Biais dans les modèles IA

Depuis 2023, l'association caritative « Open Worldwide Application Security Project » a élaboré un guide pour la sécurité des applications utilisant les modèles LLM et les modèles d'IA nommé « OWASP Top 10 for LLM applications et OWASP AI Security & Privacy guide ».

Le guide traite de:

- Risques propres aux modèles d'IA (intrusions, exfiltration, manipulations).
- Biais, la partialité et les discriminations générées par les modèles.
- Les risques systémiques (désinformation, sécurité, conformité réglementaire

❑ D’où proviennent les biais pour l’IA

Biais des données

Sous-représentation de données de cultures, de genres, de langues.

Données historiques déjà discriminantes.

Déséquilibre dans les sources d’information.

Biais de conception du modèle

Choix des paramètres d'entraînement.

Architecture favorisant certaines corrélations.

Approches d’évaluation insuffisantes.

Biais d’usage

Déploiement dans un contexte différent de celui envisagé.

Mauvaise interprétation des sorties du modèle.

Entrées manipulées par des utilisateurs malveillants (prompt injection, jailbreaking).

❑ Risques liés aux biais (OWASP TOP 10 FOR LLM)

Les biais dans les modèles de l'IA sont à la fois sources de risques éthiques et de cybersécurité.

- Les biais peuvent conduire réponses biaisées ou toxiques (**LLM07 : Insecure Output Handling**).
- récupération et réutilisation de modèles biaisés (**LLM10: Model Theft)**
- Injection volontaire de biais par des attaquants (**LLM04 : Training Data Poisoning)**

❑Approches de OWASP pour atténuer les biais et les risques y relatifs

Gouvernance des données

- Documentation complète (datasheet).
- Vérification de représentativité des datasets.
- Audit continu des sources.

Robustesse du modèle

- Tests adversariaux pour identifier des biais cachés.
- Évaluation continue par groupes de population.
- Normalisation et balancement automatique.

Monitoring post-déploiement

- Système de détection d’anomalies ou de dérive.
- Logs des interactions sensibles.
- Alertes en cas de comportements biaisés.

Transparence

- Indication claire des limites.
- Explicabilité des processus de décision.
- Inclusion d’un “AI Ethics Impact Statement”.

2. 2. Régulation de l’IA

Niveau EU AI Act (2024–2026)

Il reste un cadre structurant qui:
Obligations pour les modèles IA à risque :

- Documentation des données.
- Robustesse et cybersécurité.
- Gouvernance des biais.
- Auditabilité + enregistrement des événements.
- Transparence envers les utilisateurs.
- Interdiction de certains usages (surveillance biométrique massive, manipulation cognitive).

❑ Les modèles de fondation doivent :

- publier des résumés de données d’entraînement,
- respecter les droits d’auteur,
- intégrer des garde-fous anti-manipulations

OCDE, UNESCO, ISO

OCDE – AI Principles

Normes internationales de référence. Mots clés : équité, transparence, responsabilité, sécurité, droits humains.

UNESCO – Éthique de l’IA

Cadre global basé sur droits humains et inclusion culturelle.

ISO/IEC 42001 (2023–2025)

Gestion de la gouvernance IA (gestion des risques, conformité, qualité, traçabilité)

Alignement avec OWASP : la triade éthique

1.Sécurité

Le modèle ne doit pas causer de dommages (involontaires ou via attaques).

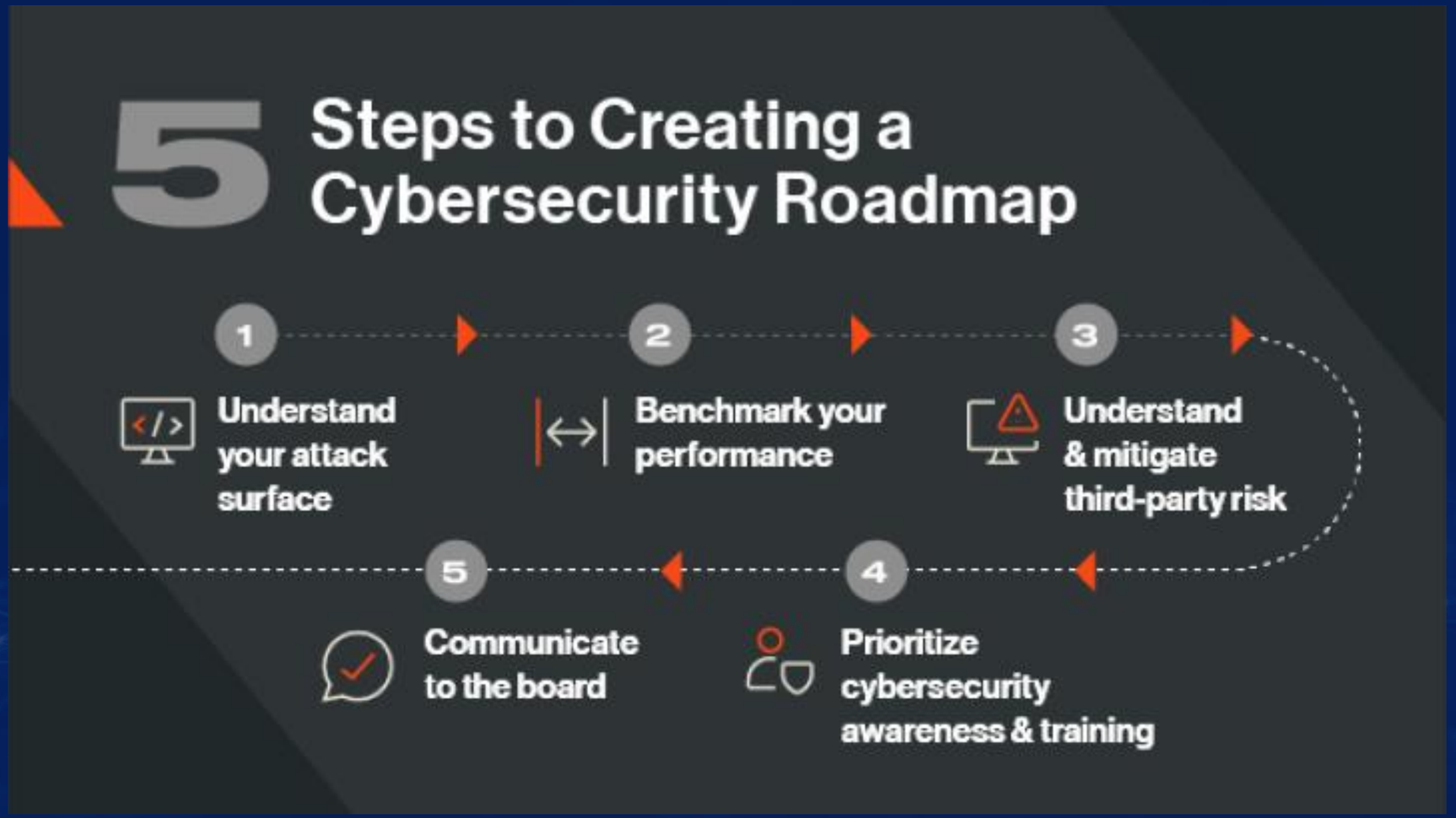
2. Robustesse

Le modèle doit fonctionner correctement et de façon équitable.

3. Transparence & responsabilité

Toute décision ou recommandation IA doit pouvoir être expliquée, tracée, auditée

••• 3. Roadmap personnel : Intégrer l'IA dans les audits sécurité



En 2026, ce dispositif modulaire doit servir de référence pour les l’ajustement des compétences.

6 à 12 mois de formation selon le niveau

Roadmap personnel est mis en place et pensé pour accompagner la formation et le recyclage des experts en sécurité:

auditeur,

pentester,

risk analyst,

Membre de SOC et SOC

Responsable de risque IA

1 Foundations : Comprendre l’IA appliquée au cyber

Objectif : acquérir les bases techniques indispensables.

Compétences à développer

- LLM, embeddings, vector stores, agents IA.
- Comprendre les risques spécifiques LLM (OWASP Top 10 LLM).
- Différents usages de l’IA en sécurité :
  - détection, corrélation d’événements, analyse de logs,
  - classification d’incidents,
  - simulation de menaces non techniques (phishing, social engineering IA).

2 Outils IA pour Auditeurs : Maîtrise pratique

IA généralistes (pour automatiser l’analyse)

- ChatGPT / Claude / Mistral / Llama
- Extensions VSCode IA pour revue de code

IA spécialisées sécurité

- Analyse de logs automatisée (Elastic + IA, Splunk AI Assist)
- Darktrace, Vectra, Microsoft Defender AI
- Analyse IA de configuration Cloud (Wiz, Lacework, Prisma, OrcaAI)
- **IA pour automatiser les tâches d’audit**
- Génération de checklist IA (ISO 27001, NIST, OWASP)
- Analyse automatique de politiques de sécurité

3 Automatisation : Construire tes assistants IA d'audit

Compétences à développer

- Prompts avancés (audit, risk, analyse de logs)
- Construction d'agents IA spécialisés : "Audit Assistant", "Log Analyzer Bot"
- Pipelines IA + outils : Python + API d'IA + SIEM/SOAR
- Détection d'anomalies sur données anonymisées
- Analyse d'architecture via IA

4 Méthodologies : IA + Cadres de conformité

Adapter l'audit à l'ère de l'IA

- Intégration IA dans un audit ISO 27001 / 27005
- Cartographie des risques IA → cybersécurité → conformité
- Audit LLM et IA selon : OWASP LLM Top 10, NIST AI RMF, EU AI Act

Projets pratiques

- Construire une **checklist IA & Sécurité** pour ton entreprise.
- Réaliser un **audit pilote IA d'une application interne**.

5 Expertise : Mesurer, Monitorer & Améliorer

Compétences clés

- Mesure continue de l'efficacité IA dans la sécurité
- Détection de dérive (model drift, data drift)
- Surveillance de l’usage interne de l’IA (Shadow AI)
- Évaluation : hallucinations, biais, exposition involontaire à des données sensibles.

Objectif final

Transformer le rôle d’auditeur en **superviseur continu des risques IA & sécurité**

6 Leadership & Impact (en continu)

Devenir un référent IA dans la sécurité

- Présenter l’approche IA lors d’audits internes.
- Former les équipes (SOC, dev, GRC).
- Mettre en place un **cadre d’usage responsable de l’IA**.
- Participer aux communautés (OWASP AI, NIST, ISO).

Ton objectif : automatiser, améliorer et fiabiliser les audits sécurité via l’IA.

Rôle final : Auditeur IA-Sécurité moderne, capable d’évaluer autant les systèmes classiques que les systèmes d’IA.

CONCLUSION

- 2026 est l'année où **l'IA devient un acteur stratégique à part entière**, capable de défendre, d'attaquer, de tromper ou de neutraliser.
- Le défi pour les organisations et les États est de garder le **contrôle humain, anticiper l'escalade et renforcer les défenses éthiques et robustes.**
- **Le défi pour les RSSI, RSI, DPO** sera de se référer des pratiques de énoncées par:
  - **OWASP:** Les biais sont un risque stratégique à traiter les vulnérabilités; ce qui met un accent sur le choix de modèles et le traitement des données d'apprentissage (ETL).
  - **OWASP recommandée :** Conception éthique dès le départ. Sécurité + intégrité des données. Évaluation indépendante. Transparence et documentation systématique. Supervision humaine
  - **AI Act, NIST2, ISO:** obligation d'audit, de traçabilité et de gouvernance.
  - **CN**: Sans être une réglementation spécifique de l'AI, demeure applicable aux projets IA

QUESTIONS / REPONSES