
OWASP
COTONOU CHAPTER

OWASP®

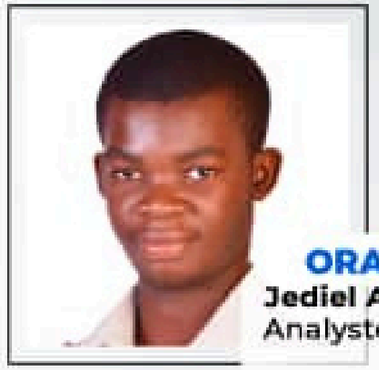
COTONOU CHAPTER

THÈME

L'IA AU SERVICE DE LA DÉTECTION DES MENACES : *Mythe Ou Réalité*

DATE	LIEU	HEURE
03 JUILLET 2025	PIGIER BÉNIN AGONTIKON	16H00

Guide de GenAI Red Teaming



ORATEUR 2

Jediel ADEFOULOU

Analyste en cybersécurité

Pour plus d'information

+229 0169292907

PLAN

01 Introduction

Le paysage des menaces évolue rapidement, et l'IA est souvent présentée comme la solution miracle.

L'IA est-elle notre sauveur face aux cybermenaces, ou une simple poudre aux yeux ?

Démystifier le rôle de l'IA dans la détection des menaces et introduire le concept crucial du "Red Teaming" pour les IA génératives.

02

Mythes vs Réalités - Démystifier l'IA en détection de menaces

Mythe I : "L'IA détecte et prévient TOUTES les menaces"

Réalité :
L'IA est un outil extrêmement puissant et un atout majeur pour la détection et la prévention des menaces, mais elle n'est ni infaillible, ni exhaustive.

- Limites techniques : Vulnérable aux attaques ciblées (data poisoning, adversarial prompts)
- Dépendance aux données : Des jeux d'entraînement obsolètes réduisent son efficacité

02

Mythes vs Réalités - Démystifier l'IA en détection de menaces

Mythe 2 : "L'IA est tellement avancée qu'elle détecte tous les nouveaux types d'attaques, même ceux qu'elle n'a jamais vus"

Réalité :

- Elle dépend toujours de ses données d'entraînement, peut repérer des comportements similaires ou légèrement modifiés, mais les attaques totalement nouvelles lui échappent.
- Les attaques ciblées, conçues spécifiquement pour une victime unique, ne sont généralement pas détectées par la seule IA

02

Mythes vs Réalités - Démystifier l'IA en détection de menaces

Mythe 3 : L'IA remplace complètement les expert humains.

Réalité :

L’IA augmente l'efficacité des analystes, notamment en triant de gros volumes de logs et en détectant des anomalies rapidement, mais elle ne remplace pas la compréhension contextuelle, la prise de décision stratégique ou l’investigation créative qu’un humain apporte.

IA manque d’intuition et de jugement pour détecter les nouvelles attaques et les attaques complexes et a besoin maintenance, tuning, supervision, et formation continue.

03

Red Teaming

Qu'est ce que c'est?

Le red teaming est une simulation d'attaque réalisée par des hackers éthiques (internes ou externes) mandatés pour imiter les tactiques réelles des cybercriminels (réseaux, applications, phishing, intrusion physique...) afin d'évaluer la sécurité globale d'une organisation.

- Il opère en aveugle, souvent sans prévenir les équipes de défense (blue team) .
- Il intègre des attentes concrètes : atteindre une base de données critique, prendre le contrôle d'un serveur .

Pourquoi ?

- Simuler des attaques réelles
- Détecter les failles cachées
- Tester les défenses en réelle situation
- Améliorer la culture sécurité

04

GenAI Red Teaming - Guide OWASP

Principaux risque à prendre en compte

- Injection rapide : tromper le modèle afin qu'il enfreigne ses règles ou divulgue des informations sensibles.
- Biais et toxicité : générer des résultats nuisibles, offensants ou injustes.
- Fuite de données : extraire des informations privées ou de la propriété intellectuelle du modèle.
- Empoisonnement des données : manipuler les données d'entraînement à partir desquelles un modèle apprend afin de le faire se comporter de manière indésirable.
- Hallucinations/confabulations : le modèle fournit avec assurance des informations erronées.
- Vulnérabilités agentives : attaques complexes contre les « agents » IA qui combinent plusieurs outils et étapes décisionnelles.

04

GenAI Red Teaming - Guide OWASP

Qu'est ce que c'est?

Le GenAI Red Teaming consiste à simuler des comportements hostiles à l'encontre de systèmes d'IA générative, tels que les grands modèles linguistiques (LLM), afin de mettre au jour les vulnérabilités liées à la sécurité, à la sûreté et à la confiance. En adoptant le point de vue d'un attaquant, nous identifions les failles avant qu'elles ne causent des dommages dans le monde réel.

Pourquoi ?

La cybersécurité traditionnelle se concentre sur les exploits techniques (par exemple, le piratage de serveurs), mais le GenAI Red Teaming examine également comment les modèles d'IA peuvent produire des résultats nuisibles ou trompeurs. Étant donné que les systèmes d'IA influencent des décisions critiques,il est essentiel de garantir leur sécurité et leur conformité avec les valeurs de l'organisation.

04

GenAI Red Teaming - Guide OWASP

Etapes

1. *Définir les objectifs et le champ d'application*
2. *Constituer l'équipe*
3. *Modélisation des menaces*
4. *Traitez l'ensemble de la pile d'applications*
5. *Utilisez des outils et des cadres*
6. *Documentez vos conclusions et établissez un rapport*
7. *Débriefing/Analyse post-engagement*
8. *Amélioration continue*

04 *Bonnes pratiques pour une IA cyber-résiliente*

Résultats

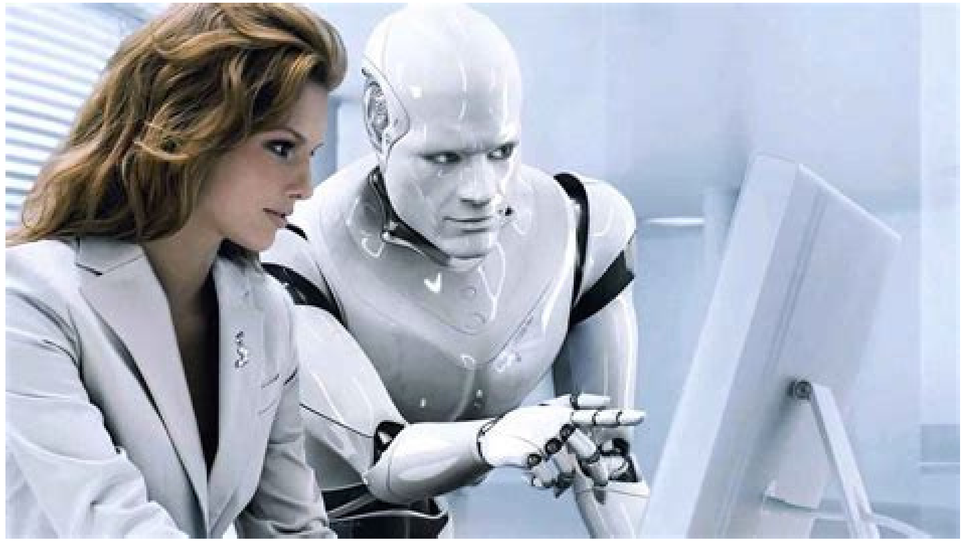
Au total on nous avons 29

- Établir des politiques, normes et procédures GenAI : Tenir compte du contexte organisationnel et créer un inventaire de l'utilisation des LLM.
- Établir des objectifs clairs : Définir des objectifs spécifiques pour chaque session de Red Teaming. 67
- Établir des critères de succès d'évaluation clairs et significatifs : Aller au-delà d'un succès/échec binaire pour les systèmes probabilistes.
- Développer des suites de tests complètes : Créer et maintenir un ensemble diversifié de cas de test et les mettre à jour régulièrement.
- Promouvoir la collaboration interfonctionnelle : Impliquer des experts de divers domaines (IA, sécurité, éthique).
- Prioriser les considérations éthiques : S'assurer que les activités de Red Teaming respectent les directives éthiques.

04

Conclusion

- Réalité : L'IA est un atout majeur pour la détection (behavioral analysis, réduction des délais), mais pas une solution autonome.
- Impératifs :
  - Utiliser le GenAI Red Teaming (Guide OWASP) pour auditer vos systèmes IA.
  - Investir dans la formation : 63% des échecs d'intégration IA sont liés à un manque de compétences 8.
- Perspective : Vers une coévolution IA / Expertise humaine face aux menaces hybrides.



04

Références

- chatgpt, gemini, deepseek
- Guide GenAI Red Teaming par OWASP
- https://hadess.io/what-is-red-team/
- https://www.bitdefender.com/en-jm/business/infozone/what-is-red-teaming

OWASP COTONOU

Merci pour votre attention !

Avez-vous des questions ?