



L'IA que vous déployez peut être attaquée.

Vous intégrez de l'IA dans vos produits, ou vos équipes s'en servent déjà tous les jours ? Vous attaquez ces systèmes vous-même - assistants, moteurs de recherche internes, agents qui ont accès à vos outils - puis vous apprenez à les refermer.

DURÉE	TARIF INTER	INTRA	NIVEAU
3 jours	2 250 € HT	dès 1 500 € HT/jour	Avancé

À l'issue de la formation

- Déployer un environnement d'IA local, souverain et reproductible pour auditer sans dépendance cloud
- Identifier et exploiter les vulnérabilités des LLM, des architectures RAG, des agents et des déploiements MCP
- Conduire un audit de sécurité d'agents IA et de serveurs MCP, de la cartographie de surface à l'impact métier
- Durcir un déploiement IA : validation entrée/sortie, guardrails, gateway de contrôle, sandboxing, moindre privilège
- Mettre en place l'observabilité et la détection des abus d'agents et d'appels de tools anormaux
- Cadrer le risque IA d'une organisation et le périmètre d'une prestation AI red team selon les référentiels OWASP

Pour qui

- Architectes et développeurs déployant des systèmes IA (RAG, agents, MCP)
- Ingénieurs AppSec et auditeurs intégrant l'IA à leur périmètre de test
- RSSI techniques souhaitant cadrer le risque IA avec une vraie compréhension du terrain
- Pentesters et red teamers qui veulent ajouter la surface IA à leur arsenal

Prérequis

Bases en Python (fichiers, API REST, subprocess), Linux et Docker. Une première expérience en sécurité ou en développement applicatif est utile. Aucune connaissance préalable en IA ou en machine learning n'est requise.

Déroulé

Jour 1 Fondations et premières attaques sur les LLM

- Panorama de la nouvelle surface : chatbots, RAG, agents, MCP, code généré par IA
- Limites fondamentales des LLM : tokens, fenêtre de contexte, prompt système, hallucinations
- Laboratoire local et souverain : Ollama, modèles open source, modèle custom, mode air-gap
- OWASP Top 10 LLM 2025 et risques Agentic AI
- Prompt injection directe et indirecte, extraction de prompt système : role-play, encodage, multi-tour, crescendo
- TP : attaque d'un chatbot à quatre niveaux de protection croissante, puis durcissement du même chatbot

Jour 2 RAG et agents : exploiter puis durcir

- Architecture RAG : embedding, base vectorielle, recherche de similarité
- Quatre surfaces RAG : poisoning, extraction, injection indirecte, manipulation de recherche
- Anatomie d'un agent : LLM, tools, mémoire, planificateur - la différence entre répondre et agir
- Surfaces agentiques : tool poisoning, manipulation de mémoire, excès d'autonomie, détournement d'objectif
- Contre-mesures : contrôle d'accès documentaire, sanitisation, moindre privilège des tools, human-in-the-loop
- TP : exploitation d'un RAG vulnérable puis de l'agent financier FinBot, avant et après garde-fous

Jour 3 MCP, durcissement et pilotage du risque

- Le protocole MCP comme surface d'attaque : serveurs, tools, ressources, chaîne d'approvisionnement
- Audit MCP offensif : tool poisoning, tool shadowing, rug pull, confused deputy sur OAuth
- Chaîne complète : prompt injection → exécution de code via MCP → vol de credentials → exfiltration cloud
- Durcir : gateway de contrôle, allow-list de tools, sandboxing, journalisation centralisée
- Observabilité et détection des abus d'agents : boucles anormales, appels de tools inattendus
- Cadrer le risque : check-list d'architecture, périmètre d'une prestation AI red team, EU AI Act, NIS2, ANSSI
- Challenge final : mini-CTF en équipes, puis challenge inversé - sécuriser un agent contre une attaque automatisée

Ce qui est compris

- Laboratoire IA local et souverain (Ollama, Docker) - aucune donnée ne sort
- Chatbot, RAG et agent volontairement vulnérables, avec leurs corrigés
- Lab MCP dédié reproduisant la chaîne complète jusqu'à l'exfiltration cloud

- Gateway défensif de référence à rejouer sur ses propres déploiements
- Check-list d'architecture et de cadrage d'un audit IA
- Environ 60 % de pratique, logique purple sur chaque module

Prochaines sessions inter-entreprises

DATES	LIEU ET FORMAT	PLACES
8 - 10 février 2027	À distance · Distanciel	7 places
7 - 9 juin 2027	Paris · Présentiel ou distanciel	8 places
18 - 20 octobre 2027	Paris · Présentiel ou distanciel	8 places

Résistez aux hackers - organisme de formation certifié Qualiopi. Prise en charge OPCO possible, les pièces sont fournies.
Inscription et devis : <https://resistez-aux-hackers.com/contact.html> - formation@resistez-aux-hackers.com
Programme susceptible d'évoluer. Version du 08/10/2026. Accessibilité : nous contacter pour toute situation de handicap, nous étudions les adaptations possibles.