Cybersécurité et IA générative

L'objectif de cette formation est de donner aux professionnels de la cybersécurité les outils pour anticiper, détecter et mitiger les vulnérabilités liées aux LLMs, à l'orchestration multi-agents, à l'usage de plugins tools et API dans les systèmes d'IA générative, et pour poser un cadre de sécurité opérationnel.

Pré-requis : *Bonnes bases en cybersécurité, connaissances générales en IA ou architectures logicielles modernes, lecture de code Python ou YAML utile (non indispensable).*

Modalité :
- Distanciel en classe virtuelle
- E-learning : à venir
- Présentiel

Communauté: community.reconvert.net

Durée totale : 21 H (3 jours)

PLAN DETAILLE

Comprendre les LLMs et les architectures agentiques

Rappels : LLMs, prompts, embeddings, context window
Architectures LangChain, CrewAI, AutoGen, Hugging Face Transformers
Chaînes, agents, tools, mémoire, exécution multi-agent

Cartographie des surfaces d'attaque dans un système LLM

Attaques sur l'entrée: prompt injection, jailbreak
Manipulation du contexte: prompt pollution, data poisoning
Tool misuse - usage malicieux d'outils exposés:
+ API non filtrée
+ exécution de code
+ accès web
Hijack d'agent ou de mémoire: state leakage, replay attack

Panorama des menaces réelles

Exemples concrets :
+ vol de données via prompt
+ désactivation de sécurité par agent
+ exfiltration d'informations
Études de cas :
+ ChatGPT
+ Plugins malveillants
+ Hallucinations exploitées
Risques directs: injection, fuite, contournement
Risques indirects: manipulation, désinformation

Sécurisation des outils, agents et chaînes

Confinement des tools: sandboxing, whitelisting, quota
Logs, audit, replay des agents
Limiter les agents autonomes : contrôle par supervision, critic agents, throttle
Safe tool composition : gestion des dépendances entre agents

Détection et monitoring en environnement LLM

Outils d'observabilité: LangSmith, TruLens, PromptLayer, Helicone
Journaux de prompt + outils utilisés + outputs générés
Détection d'anomalies dans les réponses:
+ exfiltration
+ non-conformité
+ hallucinations malicieuses

Sécuriser les entrées et les prompts

Techniques de sanitation et validation de prompt
Prompt templating sécurisé
Détection de patterns malveillants:
+ regex
+ scoring LLM
+ agents critiques)

Cybersécurité et IA générative

L'objectif de cette formation est de donner aux professionnels de la cybersécurité les outils pour anticiper, détecter et mitiger les vulnérabilités liées aux LLMs, à l'orchestration multi-agents, à l'usage de plugins tools et API dans les systèmes d'IA générative, et pour poser un cadre de sécurité opérationnel.

Pré-requis : *Bonnes bases en cybersécurité, connaissances générales en IA ou architectures logicielles modernes, lecture de code Python ou YAML utile (non indispensable).*

Modalité :
- Distanciel en classe virtuelle
- E-learning : à venir
- Présentiel

Communauté: community.reconvert.net

Durée totale : 21 H (3 jours)

PLAN DETAILLE

Atelier 1 : Risques d'un système GenAI multi-agent

Étude d'un cas concret: chatbot ou agent multi-outils
Identification des points de vulnérabilité
Proposition de sécurisation:
+ pare-feu applicatif
+ journalisation
+ isolation

Gouvernance sécurité d'un système LLM/agentic

Politiques de droits, rôles et audit
Cycle de vie sécurisé d'un prompt / modèle / outil
Revue de sécurité en phase design:
+ menace, DPIA
+ logs, validation

Sécurité dans le cycle DevSecOps IA

Intégration de la sécurité dans la CI/CD de modèles ou d'agents
Testing des prompts, scoring automatique
Méthodologie de revue de chaîne (prompt → outil → réponse)
Outils de test de vulnérabilités spécifiques aux LLMs

Cadre réglementaire et souveraineté

RGPD appliqué aux systèmes LLM (traces, personnalisation, exécution)
IA Act : contraintes sur la sécurité et la robustesse des systèmes
Hébergement sécurisé:
+ cloud souverain
+ LLM self-hosted
+ vector stores chiffrés

Atelier final : Audit d'un système LLM/Agentic

Analyse complète d'un prototype avec LangChain ou AutoGen
Identification de vulnérabilités potentielles
Élaboration d'un plan de sécurisation