Date de publication : 28/02/2025

Formation : Sécurité de l'intelligence artificielle : enjeux, risques et bonnes pratiques

Sécuriser les modèles, agents et maîtriser les risques de l'IA générative

Séminaire - 2j - 14h00 - Réf. SIA

Prix : 2170 CHF H.T.

4,4 / 5

L'IA transforme profondément les entreprises et ouvre des perspectives inédites. Mais cette révolution s'accompagne de nouveaux défis notamment en termes de sécurité. Cette formation vous propose d'identifier et comprendre les risques spécifiques liés à l'IA, tels que le prompt injection, les hallucinations, le shadow AI ou encore l'empoisonnement des données. Vous verrez comment ces menaces impactent la sécurité des systèmes IA et quelles stratégies adopter pour les anticiper, les atténuer et garantir un usage sécurisé et conforme de l'intelligence artificielle en entreprise.

Objectifs pédagogiques

À l'issue de la formation, le participant sera en mesure de :

- Identifier les nouveaux risques liés à l'intelligence artificielle
- Comprendre les vulnérabilités et attaques spécifiques à l'IA
- Maîtriser les méthodes et bonnes pratiques pour sécuriser un projet ou une application d'IA
- Exploiter l'IA générative pour renforcer la cybersécurité

Public concerné

DSI, RSSI, chefs de projets, responsables sécurité, consultants, chefs de projet IT/IA et responsable de la transformation numérique.

Prérequis

Connaissances générales des systèmes d'information.

Vérifiez que vous avez les prérequis nécessaires pour profiter pleinement de cette formation en faisant ce test.

PARTICIPANTS

DSI, RSSI, chefs de projets, responsables sécurité, consultants, chefs de projet IT/IA et responsable de la transformation numérique.

PRÉREQUIS

Connaissances générales des systèmes d'information.

COMPÉTENCES DU FORMATEUR

Les experts qui animent la formation sont des spécialistes des matières abordées. Ils ont été validés par nos équipes pédagogiques tant sur le plan des connaissances métiers que sur celui de la pédagogie, et ce pour chaque cours qu'ils enseignent. Ils ont au minimum cinq à dix années d'expérience dans leur domaine et occupent ou ont occupé des postes à responsabilité en entreprise.

MODALITÉS D'ÉVALUATION

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques...
Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

Modalités d'évaluation

Le formateur évalue la progression pédagogique du participant tout au long de la formation au moyen de QCM, mises en situation, travaux pratiques…

Le participant complète également un test de positionnement en amont et en aval pour valider les compétences acquises.

Programme de la formation

1 Sécurité de l'IA : enjeux majeurs et défis à relever

- Fondamentaux de l'IA : introduction aux concepts clés : IA générative, API, machine learning et le deep learning.
- Impact stratégique de l'IA sur les entreprises : adoption fulgurant et nouveaux défis en matière de sécurité.
- Identification des risques : exploration des menaces liées aux cyberattaques, aux biais algorithmiques et modèles d'IA.
- Approche globale de la sécurité de l'IA : dimensions techniques, juridiques, organisationnelles et éthiques.

2 Nouvelles menaces et cyberattaques

- Ingénierie sociale et fraude assistée par IA : hameçonnage par deepfake, clonage vocal, génération de sites frauduleux.
- Malware : logiciels malveillants génératifs, malwares polymorphes.
- Keylogger amélioré, obfuscation intelligente, logiciels espions furtifs, ransomware auto-évolutif.
- Attaques sur les systèmes d'authentification : contournement des systèmes biométriques, résolveurs de CAPTCHA assistés.
- Découverte automatisée de vulnérabilités, exploitation de vulnérabilités zero-day, fuzzing neuronal.
- Techniques d'évasion améliorées, attaques sur la chaîne d'approvisionnement.

3 Vulnérabilités des systèmes d'IA

- Top 10 pour LLM Applications de l'OWASP.
- Attaques sur les interactions : injection de prompts, manipulation des sorties.
- Attaques sur l'intégrité des données : empoisonnement, falsification ou corruption des jeux de données d'entraînement.
- Attaques sur la confidentialité et propriété intellectuelle : extraction ou inversion de modèle, fuite de données, etc.
- Attaques sur les mécanismes d'apprentissage et prise de décision : falsification des entrées/sorties, par bruitage, etc.
- Attaques sur les modèles : inversion, substitution, détournement, exploitation des biais, détournement des algorithmes.
- Attaques sur les infrastructures : exploitation de vulnérabilités dans les frameworks et dans les API.

MOYENS PÉDAGOGIQUES ET TECHNIQUES

• Les moyens pédagogiques et les méthodes d'enseignement utilisés sont principalement : aides audiovisuelles, documentation et support de cours, exercices pratiques d'application et corrigés des exercices pour les formations pratiques, études de cas ou présentation de cas réels pour les séminaires de formation.
• À l'issue de chaque formation ou séminaire, ORSYS fournit aux participants un questionnaire d'évaluation du cours qui est ensuite analysé par nos équipes pédagogiques.
• Une feuille d'émargement par demi-journée de présence est fournie en fin de formation ainsi qu'une attestation de fin de formation si le participant a bien assisté à la totalité de la session.

MODALITÉS ET DÉLAIS D'ACCÈS

L'inscription doit être finalisée 24 heures avant le début de la formation.

ACCESSIBILITÉ AUX PERSONNES HANDICAPÉES

Pour toute question ou besoin relatif à l'accessibilité, vous pouvez joindre notre équipe PSH par e-mail à l'adresse psh-accueil@orsys.fr.

4 Gestion des risques dans les projets d'intelligence artificielle

- Cartographie des risques du MIT : AI Risk Repository.
- Taxonomie des menaces : MITRE ATLAS, OWASP AI Exchange et NIST Adversarial ML.
- Référentiels de gestion des risques : NIST AI Risk Management Framework (RMF), méthodologie EBIOS RM.
- La gestion des risques dans la mise en œuvre d'un SMIA selon l'ISO 42001.
- Protection des données personnelles et évaluation d'impact sur la vie privée (PIA/AIPD).
- Traitement des risques : mesures techniques et organisationnelles.

5 Sécurité des applications IA (LLM, agents et APIs)

- A : Principes généraux de sécurité des applications d'IA :
- Approche Secure AI by design.
- Identification de la surface d'attaque et des menaces spécifiques aux modèles IA.
- Contrôle d'accès et chiffrement des données.
- B : Sécurité des modèles et des algorithmes :
- Sécurité des datasets (protection des données d'entraînement, prévention du data poisoning et des fuites de données).
- Hardening des modèles d'IA : adversarial training, model watermarking.
- Techniques de détection des biais et dérives des modèles IA.
- Protection contre le prompt injection et les attaques adversariales sur les LLMs.
- C : Sécurité des infrastructures et des APIs :
- Sécurisation des API (authentification, contrôle des accès, validation des entrées/sorties).
- Sécurisation des pipelines MLOps et LLMOps : gestion des accès, intégrité des modèles.
- Validation et surveillance des modèles en production, détection d'anomalies.

6 Sécuriser l'usage de l'IA générative en entreprise

- A : Identifier et comprendre les risques liés à l'usage de l'IA générative :
- Fuites de données sensibles : exposition involontaire d'informations internes, personnelles et secrets industriels.
- Génération de réponses incorrectes ou trompeuses pouvant impacter la prise de décision.
- Shadow AI et usages non maîtrisés : déploiement incontrôlé d'outils IA en dehors du cadre autorisé par l'entreprise.
- Propriété intellectuelle et cadre juridique : responsabilités associées aux contenus générés par l'IA.
- B : Définition d'une politique d'usage des IA génératives en entreprise :
- Rédaction d'une charte d'utilisation interne des solutions d'IAGen.
- Délimitation des cas d'usage autorisés et interdits.
- Sensibilisation et formation des collaborateurs aux bonnes pratiques.
- C : Sensibilisation et gouvernance de l'usage de l'IA en entreprise :
- Formation continue des employés sur les risques et bonnes pratiques.
- Désignation d'un responsable IA (Chief AI Officer - CAIO) ou d'un comité de gouvernance IA.
- Mise en place d'un suivi et d'un reporting sur l'utilisation de l'IA dans l'entreprise.
- D : Bonnes pratiques pour un usage sécurisé :
- Limitation des prises de décision basées uniquement sur des recommandations générées.

7 Audit, transparence et évaluation de la sécurité des systèmes IA

- A : Conformité et audits des systèmes IA :
- Évaluation des risques IA avec COMPL-AI, AI Risk Repository (MIT).
- Alignement avec les exigences de l'ISO 42001 (SMIA).
- Politiques de transparence et documentation des décisions IA.
- B : Transparence et explicabilité des modèles IA :
- Importance de la transparence et de l'auditabilité des modèles IA (XAI - Explainable AI).
- Outils et méthodologies pour auditer un modèle IA (explicabilité, robustesse, dérives).
- Validation de l'équité des modèles IA et gestion des dérives dans le temps.

8 Exploiter l'IA générative dans la cybersécurité

- A : Gouvernance et de la gestion des risques (GRC) :
- Automatisation des tâches stratégiques : assistance à la rédaction des PSSI, synthèse des réglementations en vigueur.
- Veille proactive : analyse et structuration des informations issues des évolutions réglementaires.
- B : Support au développement et à la sécurisation des applications :
- Automatisation des tests de sécurité : création de scénarios d'attaques et génération de payloads polymorphiques.
- C : Renforcement des capacités de détection et de réponse aux incidents :
- Analyse avancée des logs et reconstruction des incidents et Génération de plans de réponse automatisés.
- D : Cybersurveillance et renseignement sur les menaces (Threat Intelligence) :
- Analyse et contextualisation des cybermenaces : traduction automatique et catégorisation des attaques émergentes.
- E : Assistance à la conformité réglementaire :
- Analyse des écarts et conformité aux normes : interprétation des exigences des cadres réglementaires (RGPD, DORA, NIS2).
- F : Intégration de l'IA dans les solutions de cybersécurité :
- Pare-feu XDR, SOAR, Microsoft Security Copilot, IBM Watson GenAI, Darktrace Prevent/Antigena, Zynamp, Cymulate.

Dates et lieux

CLASSE À DISTANCE
2026 : 18 juin, 24 sep., 1 déc.