Publication date : 02/28/2025

Course : Artificial intelligence security: challenges, risks and best practices

Securing models, agents and managing the risks of generative AI

Seminar - 2d - 14h00 - Ref. SIA

Price : 2170 CHF E.T.

4,4 / 5

AI is profoundly transforming businesses and opening up unprecedented prospects. But this revolution also brings new challenges, particularly in terms of security. This training course will help you identify and understand the specific risks associated with AI, such as prompt injection, hallucinations, shadow AI and data poisoning. You'll see how these threats impact the security of AI systems, and what strategies you can adopt to anticipate and mitigate them, and ensure the secure and compliant use of artificial intelligence in the enterprise.

Teaching objectives

At the end of the training, the participant will be able to:

- Identifying new risks linked to artificial intelligence
- Understanding AI-specific vulnerabilities and attacks
- Master the methods and best practices for securing an AI project or application
- Harnessing generative AI to strengthen cybersecurity

Intended audience

CIOs, CISOs, project managers, security managers, consultants, IT/IA project managers and digital transformation managers.

Prerequisites

General knowledge of information systems.

Course schedule

PARTICIPANTS

CIOs, CISOs, project managers, security managers, consultants, IT/IA project managers and digital transformation managers.

PREREQUISITES

General knowledge of information systems.

TRAINER QUALIFICATIONS

The experts leading the training are specialists in the covered subjects. They have been approved by our instructional teams for both their professional knowledge and their teaching ability, for each course they teach. They have at least five to ten years of experience in their field and hold (or have held) decision-making positions in companies.

ASSESSMENT TERMS

The trainer evaluates each participant's academic progress throughout the training using multiple choice, scenarios, hands-on work and more.
Participants also complete a placement test before and after the course to measure the skills they've developed.

1 AI security: major issues and challenges

- AI fundamentals: introduction to key concepts: generative AI, APIs, machine learning and deep learning.
- AI's strategic impact on business: meteoric adoption and new security challenges.
- Risk identification: exploring the threats posed by cyber attacks, algorithmic biases and AI models.
- Comprehensive approach to AI security: technical, legal, organizational and ethical dimensions.

2 New threats and cyber attacks

- AI-assisted social engineering and fraud: deepfake phishing, voice cloning, fraudulent site generation.
- Malware: generative malware, polymorphic malware.
- Enhanced keylogger, intelligent obfuscation, stealth spyware, self-evolving ransomware.
- Attacks on authentication systems: bypassing biometric systems, assisted CAPTCHA resolvers.
- Automated vulnerability discovery, zero-day vulnerability exploitation, neural fuzzing.
- Improved evasion techniques, supply chain attacks.

3 AI system vulnerabilities

- Top 10 for LLM Applications from OWASP.
- Attacks on interactions: injection of prompts, output manipulation.
- Attacks on data integrity: poisoning, falsification or corruption of training data sets.
- Attacks on confidentiality and intellectual property: model extraction or inversion, data leakage, etc.
- Attacks on learning and decision-making mechanisms: falsification of inputs/outputs, noise, etc.
- Attacks on models: inversion, substitution, hijacking, exploitation of biases, hijacking of algorithms.
- Attacks on infrastructures: exploiting vulnerabilities in frameworks and APIs.

4 Risk management in artificial intelligence projects

- MIT risk mapping: AI Risk Repository.
- Threat taxonomy: MITRE ATLAS, OWASP AI Exchange and NIST Adversarial ML.
- Risk management frameworks: NIST AI Risk Management Framework (RMF), EBIOS RM methodology.
- Risk management in the implementation of an ISO 42001 SMIA.
- Personal data protection and privacy impact assessment (PIA/AIPD).
- Risk management: technical and organizational measures.

TEACHING AIDS AND TECHNICAL RESOURCES

• The main teaching aids and instructional methods used in the training are audiovisual aids, documentation and course material, hands-on application exercises and corrected exercises for practical training courses, case studies and coverage of real cases for training seminars.
• At the end of each course or seminar, ORSYS provides participants with a course evaluation questionnaire that is analysed by our instructional teams.
• A check-in sheet for each half-day of attendance is provided at the end of the training, along with a course completion certificate if the trainee attended the entire session.

TERMS AND DEADLINES

Registration must be completed 24 hours before the start of the training.

ACCESSIBILITY FOR PEOPLE WITH DISABILITIES

Do you need special accessibility accommodations? Contact Mrs. Fosse, Disability Manager, at psh-accueil@orsys.fr to review your request and its feasibility.

5 AI application security (LLM, agents and APIs)

- A: General safety principles for AI applications :
- Secure AI by design approach.
- Identify the attack surface and threats specific to AI models.
- Access control and data encryption.
- B: Model and algorithm security :
- Dataset security (protection of training data, prevention of data poisoning and data leakage).
- Hardening AI models: adversarial training, model watermarking.
- Techniques for detecting bias and drift in AI models.
- Protection against prompt injection and adversarial attacks on LLMs.
- C: Infrastructure and API security :
- API security (authentication, access control, input/output validation).
- Securing MLOps and LLMOps pipelines: access management, model integrity.
- Validation and monitoring of models in production, detection of anomalies.

6 Securing the use of generative AI in business

- A: Identify and understand the risks associated with the use of generative AI :
- Sensitive data leaks: inadvertent exposure of internal, personal information and industrial secrets.
- Generation of incorrect or misleading answers that can impact decision-making.
- Shadow AI and uncontrolled use: uncontrolled deployment of AI tools outside the framework authorized by the company.
- Intellectual property and legal framework: responsibilities associated with AI-generated content.
- B: Defining a policy for the use of generative AI in companies :
- Drafting of a charter for internal use of IAGen solutions.
- Delimitation of authorized and prohibited uses.
- Raising awareness and training employees in best practices.
- C: Awareness and governance of AI use in business :
- Ongoing employee training on risks and best practices.
- Appointment of a Chief AI Officer (CAIO) or an AI governance committee.
- Set up monitoring and reporting on the use of AI in the company.
- D: Best practices for safe use :
- Limit decision-making based solely on generated recommendations.

7 Audit, transparency and safety assessment of AI systems

- A: Compliance and audits of IA systems :
- AI risk assessment with COMPL-AI, AI Risk Repository (MIT).
- Alignment with ISO 42001 (SMIA) requirements.
- Transparency policies and documentation of IA decisions.
- B: Transparency and explicability of AI models :
- Importance of transparency and auditability of AI models (XAI - Explainable AI).
- Tools and methodologies for auditing an AI model (explicability, robustness, drifts).
- Validation of the fairness of AI models and management of drifts over time.

8 Harnessing generative AI for cybersecurity

- A: Governance and risk management (GRC) :
- Automation of strategic tasks: assistance with the drafting of PSSI, summary of current regulations.
- Proactive intelligence: analysis and structuring of information arising from regulatory developments.
- B: Application development and security support :
- Automated security testing: creation of attack scenarios and generation of polymorphic payloads.
- C: Strengthening incident detection and response capabilities :
- Advanced log analysis, incident reconstruction and automated response plan generation.
- D: Cybersurveillance and Threat Intelligence :
- Analysis and contextualization of cyberthreats: automatic translation and categorization of emerging attacks.
- E: Regulatory compliance assistance :
- Gap analysis and standards compliance: interpreting the requirements of regulatory frameworks (RGPD, DORA, NIS2).
- F: Integrating AI into cybersecurity solutions:
- XDR Firewall, SOAR, Microsoft Security Copilot, IBM Watson GenAI, Darktrace Prevent/Antigena, Zynamp, Cymulate.

Dates and locations

REMOTE CLASS
2026 : 18 June, 24 Sep., 1 Dec.