Publicatiedatum : 28/02/2025

Opleiding : Beveiliging van kunstmatige intelligentie: problemen, risico's en best practices

Modellen, agenten en de risico's van generatieve AI beveiligen

seminarie - 2d - 14u00 - Ref. SIA

Prijs : 1850 € V.B.

4,4 / 5

AI verandert bedrijven ingrijpend en biedt ongekende mogelijkheden. Deze revolutie brengt echter ook nieuwe uitdagingen met zich mee, met name op het gebied van beveiliging. Deze cursus helpt je de specifieke risico's van AI te identificeren en te begrijpen, zoals prompt injection, hallucinaties, schaduw-AI en datavergiftiging. U zult zien hoe deze bedreigingen de beveiliging van AI-systemen beïnvloeden en welke strategieën u moet toepassen om hierop te anticiperen en ze te beperken, en om een veilig en compliant gebruik van kunstmatige intelligentie in het bedrijfsleven te garanderen.

Pedagogische doelstellingen

Aan het einde van de training is de deelnemer in staat om:

- De nieuwe risico's van kunstmatige intelligentie identificeren
- AI-specifieke kwetsbaarheden en aanvallen begrijpen
- Beheers de methoden en best practices voor het beveiligen van een AI-project of -toepassing
- Generatieve AI inzetten om de cyberbeveiliging te verbeteren

Doelgroep

CIO's, CISO's, projectmanagers, beveiligingsmanagers, consultants, IT/IA-projectmanagers en digitale transformatiemanagers.

Voorafgaande vereisten

Algemene kennis van informatiesystemen.

Opleidingsprogramma

DEELNEMERS

CIO's, CISO's, projectmanagers, beveiligingsmanagers, consultants, IT/IA-projectmanagers en digitale transformatiemanagers.

VOORAFGAANDE VEREISTEN

Algemene kennis van informatiesystemen.

VAARDIGHEDEN VAN DE CURSUSLEIDER

De deskundigen die de cursus leiden zijn specialisten op het betreffende vakgebied. Zij werden geselecteerd door onze pedagogische teams zowel om hun vakkennis als hun pedagogische vaardigheden voor elke cursus die zij geven. Zij hebben minstens vijf tot tien jaar ervaring in hun vakgebied en oefenen of oefenden verantwoordelijke bedrijfsfuncties uit.

BEOORDELINGSMODALITEITEN

De cursusleider beoordeelt de pedagogische vooruitgang van de deelnemer gedurende de gehele cursus aan de hand van meerkeuzevragen, praktijksituaties, praktische opdrachten, ...
De deelnemer legt ook van tevoren en naderhand een test af ter bevestiging van de verworven kennis.

1 AI-beveiliging: belangrijke kwesties en uitdagingen voor de toekomst

- Grondbeginselen van AI: inleiding tot belangrijke concepten: generatieve AI, API's, machinaal leren en deep learning.
- De strategische impact van AI op bedrijven: razendsnelle adoptie en nieuwe beveiligingsuitdagingen.
- Risico-identificatie: het verkennen van bedreigingen door cyberaanvallen, algoritmische vertekeningen en AI-modellen.
- Wereldwijde aanpak van AI-beveiliging: technische, juridische, organisatorische en ethische dimensies.

2 Nieuwe bedreigingen en cyberaanvallen

- Social engineering en AI-ondersteunde fraude: deepfake phishing, stemmen klonen, frauduleuze sites genereren.
- Malware: generatieve malware, polymorfe malware.
- Verbeterde keylogger, intelligente verduistering, stealth spyware, zelfontwikkelende ransomware.
- Aanvallen op verificatiesystemen: omzeilen van biometrische systemen, geassisteerde CAPTCHA-resolvers.
- Geautomatiseerde ontdekking van kwetsbaarheden, exploitatie van 'zero-day'-kwetsbaarheden, neural fuzzing.
- Verbeterde ontwijkingstechnieken, aanvallen op de toeleveringsketen.

3 Kwetsbaarheden in AI-systemen

- Top 10 voor LLM-sollicitaties van OWASP.
- Aanvallen op interacties: injectie van prompts, manipulatie van uitvoer.
- Aanvallen op gegevensintegriteit: vergiftiging, vervalsing of corruptie van trainingsdatasets.
- Aanvallen op vertrouwelijkheid en intellectueel eigendom: modelextractie of inversie, lekken van gegevens, enz.
- Aanvallen op leer- en besluitvormingsmechanismen: vervalsing van input/output, door ruis, enz.
- Aanvallen op modellen: inversie, substitutie, kaping, exploitatie van vertekeningen, kaping van algoritmen.
- Aanvallen op infrastructuren: kwetsbaarheden in frameworks en API's uitbuiten.

4 Risicobeheer in projecten voor kunstmatige intelligentie

- MIT-risicokartering: AI-risicorepository.
- Dreigingstaxonomie: MITRE ATLAS, OWASP AI Exchange en NIST Adversarial ML.
- Risicobeheerraamwerken: NIST AI Risk Management Framework (RMF), EBIOS RM-methodologie.
- Risicobeheer bij de implementatie van een ISMS volgens ISO 42001.
- Bescherming van persoonsgegevens en privacyeffectbeoordeling (PIA/AIPD).
- Risicobeheer: technische en organisatorische maatregelen.

PEDAGOGISCHE EN TECHNISCHE MIDDELEN

• De gebruikte pedagogische middelen en cursusmethoden zijn voornamelijk: audiovisuele hulpmiddelen, documentatie en cursusmateriaal, praktische oefeningen en correcties van de oefeningen voor praktijkstages, casestudies of reële voorbeelden voor de seminars.

• Na afloop van de stages of seminars verstrekt ORSYS de deelnemers een evaluatievragenlijst over de cursus die vervolgens door onze pedagogische teams wordt geanalyseerd.

• Na afloop van de cursus wordt een presentielijst per halve dag verstrekt, evenals een verklaring van de afronding van de cursus indien de stagiair alle sessies heeft bijgewoond.

TOEGANGSMODALITEITEN EN TERMIJNEN

De inschrijving dient 24 uur voor aanvang van de cursus plaatsgevonden te hebben.

TOEGANKELIJKHEID VOOR MINDERVALIDEN

Is voor u speciale toegankelijkheid vereist? Neem contact op met mevr. FOSSE, contactpersoon voor mindervaliden, via het adres psh-accueil@ORSYS.fr om uw verzoek en de haalbaarheid daarvan zo goed mogelijk te bestuderen.

5 Beveiliging van AI-toepassingen (LLM, agents en API's)

- A: Algemene veiligheidsprincipes voor AI-toepassingen :
- Veilige AI door ontwerpbenadering.
- Identificatie van het aanvalsoppervlak en bedreigingen die specifiek zijn voor AI-modellen.
- Toegangscontrole en gegevensversleuteling.
- B: Beveiliging van modellen en algoritmen :
- Beveiliging van datasets (bescherming van trainingsgegevens, voorkomen van datapoisoning en datalekken).
- AI-modellen verharden: adversaire training, modelwatermerken.
- Technieken voor het detecteren van bias en drift in AI-modellen.
- Bescherming tegen promptinjectie en aanvallen van tegenstanders op LLM's.
- C: Infrastructuur en API-beveiliging :
- API-beveiliging (authenticatie, toegangscontrole, validatie van invoer/uitvoer).
- Beveiligen van MLOps en LLMOps pijplijnen: toegangsbeheer, modelintegriteit.
- Validatie en bewaking van modellen in productie, detectie van afwijkingen.

6 Het gebruik van generatieve AI in het bedrijfsleven veiligstellen

- A: De risico's identificeren en begrijpen die gepaard gaan met het gebruik van generatieve AI :
- Lekken van gevoelige gegevens: onbedoelde blootstelling van interne en persoonlijke informatie en bedrijfsgeheimen.
- Het genereren van onjuiste of misleidende antwoorden die de besluitvorming kunnen beïnvloeden.
- Schaduw-AI en ongecontroleerd gebruik: ongecontroleerde inzet van AI-tools buiten het door het bedrijf toegestane kader.
- Intellectueel eigendom en het rechtskader: verantwoordelijkheden in verband met AI-gegenereerde inhoud.
- B: Definitie van een beleid voor het gebruik van generatieve AI in bedrijven :
- Opstellen van een intern gebruikershandvest voor IAGen-oplossingen.
- Afbakening van toegestaan en verboden gebruik.
- Bewustmaking en training van werknemers in goede praktijken.
- C: Bewustzijn en governance van het gebruik van AI in bedrijven :
- Voortdurende training van werknemers over risico's en goede praktijken.
- Aanstelling van een Chief AI Officer (CAIO) of een AI-governancecommissie.
- Monitoring en rapportage opzetten over het gebruik van AI in het bedrijf.
- D: Goede praktijken voor veilig gebruik :
- De besluitvorming beperken tot alleen gegenereerde aanbevelingen.

7 Audit, transparantie en veiligheidsbeoordeling van AI-systemen

- A: Naleving en audits van IA-systemen :
- AI-risicobeoordeling met COMPL-AI, AI Risk Repository (MIT).
- Afstemming op de vereisten van ISO 42001 (SMIA).
- Transparantiebeleid en documentatie van IA-besluiten.
- B: Transparantie en verklaarbaarheid van AI-modellen :
- Het belang van transparantie en controleerbaarheid van AI-modellen (XAI - Explainable AI).
- Hulpmiddelen en methodologieën voor het controleren van een AI-model (verklaarbaarheid, robuustheid, afwijkingen).
- Validatie van de eerlijkheid van AI-modellen en beheer van drift in de tijd.

8 Generatieve AI inzetten voor cyberbeveiliging

- A: Governance en risicobeheer (GRC) :
- Automatisering van strategische taken: hulp bij het opstellen van PSSI, overzicht van geldende voorschriften.
- Proactieve controle: analyse en structurering van informatie die voortkomt uit ontwikkelingen op het gebied van regelgeving.
- B: Applicatieontwikkeling en beveiligingsondersteuning :
- Geautomatiseerde beveiligingstesten: aanvalsscenario's maken en polymorfe payloads genereren.
- C: Versterking van de mogelijkheden voor opsporing van en respons op incidenten :
- Geavanceerde analyse van logs en reconstructie van incidenten en het genereren van geautomatiseerde reactieplannen.
- D: Cybersurveillance en informatie over bedreigingen :
- Analyse en contextualisering van cyberbedreigingen: automatische vertaling en categorisering van opkomende aanvallen.
- E: Hulp bij naleving van regelgeving :
- Analyse van hiaten en naleving van normen: de vereisten van regelgevingskaders interpreteren (RGPD, DORA, NIS2).
- F: AI integreren in cyberbeveiligingsoplossingen :
- XDR firewall, SOAR, Microsoft Security Copilot, IBM Watson GenAI, Darktrace Prevent/Antigena, Zynamp, Cymulate.

Data en plaats

KLAS OP AFSTAND
2026 : 18 juni, 24 sep., 1 dec.

PARIS LA DÉFENSE
2026 : 4 juni, 22 sep., 26 nov.