COURSE SYLLABUS

Securing AI/ML Systems: Protecting the Intelligent Attack Surface (Self-Paced)

Self-paced AI/ML security across ten modules: adversarial machine learning, generative AI and prompt injection, model integrity, supply chain, and pipeline security.

Group classes (Live Online) and onsite training are available for this course. For more information, email corporate@nobledesktop.com or visit: https://www.nobledesktop.com/classes/securing-ai-ml-systems-self-paced



hello@nobledesktop.com • (212) 226-4149

Course Outline

Module 1

- Know What You're Defending: AI/ML Security Foundations
- AI, machine learning, deep learning, generative AI and foundation-model concepts
- System architecture components
- Assets, trust boundaries, interfaces and dependencies
- Traditional versus AI-specific risk
- Security, resilience, robustness, privacy and trustworthiness

Module 2

- Follow the Model: The AI/ML Security Lifecycle
- Security across design, development, training, testing, deployment, operation and retirement
- Responsibilities across developers, operators, users and service providers
- Risks in experimentation and model development environments
- Protecting development, testing, staging and production
- Security gates through the lifecycle

Module 3

- Data Is the Fuel: Securing Training and Operational Data
- Training, validation, testing, inference and operational datasets
- Confidentiality, integrity, availability and provenance
- Access control for sensitive datasets
- Data poisoning and manipulation
- Protecting collection, labeling, transformation and preprocessing
- Monitoring quality, integrity and lineage

Module 4

- Attack the Learning Process: Adversarial Machine Learning
- Attacker objectives
- Poisoning attacks against training data and learning processes
- Evasion attacks against deployed models
- Privacy attacks targeting models and training information
- Model extraction and information disclosure
- Defensive concepts for robustness and resilience

Module 5

- Guard the Model: Model Integrity & Intellectual Property
- Models, weights, parameters and configurations as protected assets
- Securing repositories and storage
- Access control for model files, checkpoints and artifacts
- Provenance and integrity verification before deployment
- Theft, substitution and tampering
- Secure versioning, approval and change management

Module 6

- Secure the AI Supply Chain: Models, Libraries & Dependencies
- Third-party models, datasets, libraries, frameworks and services as dependencies
- Evaluating externally sourced components
- Verifying provenance and integrity of acquired artifacts
- Malicious or vulnerable dependencies in development environments
- Controlling imports, updates, plugins and integrations
- Ongoing supplier monitoring

Module 7

- When AI Starts Talking: Generative AI & LLM Security
- Security implications of large language models
- Direct and indirect prompt injection
- Sensitive information disclosure and unintended exposure
- Insecure output handling and downstream risk
- Retrieval-augmented generation, plugins, tools and external data sources
- Isolation, access control, validation and least privilege for generative AI applications

Module 8

- Secure the Pipeline: MLOps, APIs & Deployment
- Protecting pipelines, orchestration platforms and automation workflows
- Securing repositories, build environments and deployment
- Protecting secrets, keys, tokens and credentials
- Authentication and authorization for AI services and APIs
- Separating development, training, testing and production privileges
- Detecting unauthorized change

Module 9

- Watch the Machine: Monitoring, Detection & Incident Response

- Security logging and telemetry for AI environments
- Monitoring model access, administrative actions, API activity and configuration change
- Detecting abnormal inputs, outputs, behavior and usage
- Distinguishing performance degradation from compromise
- AI-specific incident response
- Containment, model rollback, recovery and post-incident validation

Module 10

- Govern, Test, Defend: AI Security Risk Workshop
- Identify assets, trust boundaries and dependencies in a fictional AI system, develop AI-specific threat scenarios and attack paths, find data, model, pipeline, infrastructure and generative AI vulnerabilities, evaluate likelihood and impact, then select, prioritize and defend a risk-based hardening plan