AI for Cybersecurity

Course Overview

Artificial Intelligence (AI) systems introduce new security risks that traditional application security practices weren't designed to address. Large Language Models (LLMs), AI agents, and automated decision systems are inherently non-deterministic, operate on untrusted inputs, and interact with external tools and data sources. Securing these systems requires new skills, patterns, and evaluation approaches.

The AI for Secure & Trustworthy AI Systems course prepares learners to design, build, and operate AI-enabled applications securely. The course focuses on real-world security risks such as prompt injection, unsafe context handling, over-permissioned agents, and operational failures in AI-assisted development and incident response.

Rather than focusing on theory alone, the course emphasizes applied security practices. Learners work with realistic scenarios, tooling, and workflows to secure AI inputs, harden agent integrations, operationalize AI in DevSecOps pipelines, and conduct AI-assisted incident response.

By the end of the course, learners will have designed and demonstrated a secure, production-ready AI-enabled application or workflow that addresses common AI security threats.

Who This Course Is For

This course is designed for learners who want practical experience securing AI-enabled systems, including:

- Application security engineers
- DevSecOps and platform engineers
- Software engineers working with LLMs or AI agents
- Security analysts and incident responders
- Technical professionals responsible for deploying AI in production environments

A basic familiarity in software development and security concepts is helpful, but the course is structured to guide learners step-by-step through increasingly complex AI security challenges.

What You Will Learn

By completing this course, you will be able to:

- Explain how AI systems introduce new security risks
- Secure LLM inputs, prompts, and external context
- Defend against direct and indirect prompt injection attacks
- Harden AI agent integrations using least-privilege principles
- Prevent unauthorized tool use and privilege escalation

- Integrate AI securely into DevSecOps pipelines
- Validate AI-generated code and security recommendations
- Use AI to assist with incident response and log analysis
- Evaluate AI system behaviour for safety, reliability, and misuse

How the Course Works

- **Duration:** 8 weeks (24 total instructional hours)
- **Format:** Instructor-led sessions with guided labs
- **Learning Style:** Applied, skills-based, and project-focused

Each week includes:

- A focused lecture introducing core concepts
- A hands-on lab using industry-standard tools
- A short knowledge check or take-home assignment

From **Week 4 through Week 8**, learners work on a cumulative **Capstone Project** that integrates skills from across the course.

Weekly Course Outline

Week	Topics	Overview	Focus Areas
1	Foundations of AI Security	Learn why traditional security models fall short for AI systems. Explore how non-deterministic behaviour, untrusted inputs, and external context introduce new attack surfaces	• AI threat models • Non-determinism • Adapting security frameworks
2	Secure LLM Inputs and Context	Learn how to protect AI applications from prompt injection and untrusted external data. Practice input validation, context isolation, and defensive prompt engineering.	• Direct and indirect prompt injection • Retrieval-Augmented Generation (RAG) poisoning • Input sanitization
3	Hardening AI Agent Integrations	Explore how AI agents interact with tools and APIs. Learn to prevent the “confused deputy” problem, scope permissions, and enforce least privilege.	• Agent security patterns • Application Programming Interface (API) gateways • Authorization controls
4	Security, Architecture & Capstone Proposal	Learn how to design secure AI system architectures. Begin the Capstone Project by proposing a secure AI-enabled application or workflow.	• Secure design • Threat modelling • Project planning
5	Operationalizing AI in DevSecOps	Integrate AI into Continuous Integration and Continuous Delivery/Deployment (CI/CD) pipelines for automated code	• CI/CD integration • Static Application Security Testing (SAST) tools

Week	Topics	Overview	Focus Areas
		review and vulnerability remediation. Learn how to validate AI-generated fixes safely.	• Validating AI outputs
6	Managing AI Risk in Development Pipelines	Address false positives, false negatives, and hallucinated security advice. Learn how to balance automation with human oversight.	• Evaluation strategies • Reliability • Risk management
7	AI-Assisted Incident Response	Use AI to assist with log analysis, forensics, and timeline reconstruction. Practice formulating precise queries to avoid ambiguity and missed evidence.	• Log analysis • Indicators of Compromise (IOCs) • AI-assisted investigation
8	Secure AI Operations & Capstone Completion	Evaluate AI systems in production and complete the Capstone Project. Present a secure, end-to-end AI workflow with clear safeguards and evaluation criteria.	• Operational security • Observability • System evaluation

Capstone Project: Secure AI Application or Workflow

The **Capstone Project** is the core applied component of the course. Learners will design and demonstrate a secure AI-enabled system that:

- Accepts and validates untrusted inputs safely
- Prevents prompt injection and unsafe context usage
- Restricts agent actions to authorized operations
- Integrates AI securely into development workflows
- Uses AI responsibly during incident response

Learners are not required to build a custom AI model. The emphasis is on secure integration, risk mitigation, and operational reliability.

The capstone is designed to be:

- Practical and job-relevant
- Portfolio-ready
- A clear demonstration of applied AI security skills

Assessment & Completion

This course uses a **Complete / Incomplete** model focused on demonstrated skills rather than exams.

To successfully complete the course, learners must:

- Participate in weekly labs and activities

- Complete required assignments and knowledge checks
- Contribute to and present the Capstone Project

Completion indicates readiness to work with real-world data engineering systems.

What You'll Leave With

By the end of the course, you will have:

- Hands-on experience securing AI applications
- Practical knowledge of AI-specific security threats
- A secure, end-to-end AI project you can explain and defend
- Skills applicable to production AI and security roles portfolio