AI-300

OSAI+

OffSec AI Red Teamer

Advanced AI Red Teaming (AI-300) is OffSec's hands-on training program for security professionals looking to assess and exploit modern AI-enabled systems. The course teaches learners how to identify and exploit vulnerabilities across generative AI applications, AI agents, machine learning pipelines, and supporting infrastructure. Emphasizing practical, lab-driven learning, AI-300 develops the offensive skills and adversary mindset required to test real-world AI environments and uncover emerging security risks.

Introduction to Red Teaming AI Systems	Understand how artificial intelligence systems change the traditional attack surface. This module introduces the core concepts of AI cybersecurity, explains how adversaries target AI-enabled environments, and maps AI attacks to the red team lifecycle
Reconnaissance for AI Targets	Learn how to identify and map AI applications, machine learning components, and model infrastructure within a target environment. Students practice reconnaissance techniques used to discover AI assets, dependencies, and exposed services without alerting defenders
Attacking AI Agents	Explore offensive techniques for manipulating AI agents by abusing prompt instructions, memory systems, and tool integrations. This module demonstrates how attackers influence autonomous AI applications while maintaining stealth
Attacking Multi-Agent Systems and A2A Protocols	Analyze the architecture of multi-agent AI systems and learn how adversaries exploit trust relationships between agents. Students practice attacks such as message manipulation, agent impersonation, and workflow corruption
Exploiting RAG Pipelines	Examine how attackers compromise retrieval-augmented generation (RAG) systems by poisoning knowledge sources and manipulating retrieval layers to control model outputs
Attacking Embeddings	Understand the role of embeddings in machine learning systems and perform attacks such as embedding inversion and information extraction to recover sensitive data from AI models
Attacking Model Context Protocol and Tool Surfaces	Explore how orchestration layers and AI tool integration frameworks can be abused to escalate privileges or execute unintended actions within AI systems
AI Supply Chain Attacks	Learn how adversaries target the AI supply chain, including datasets, model weights, adapters, and dependencies. Students practice techniques used to introduce malicious artifacts into AI environments before deployment
AI Infrastructure and Deployment Exploits	Identify vulnerabilities in AI infrastructure, including cloud AI platforms, model servers, and containerized machine learning workloads

Threat Modeling for AI-Enabled Targets	Develop strategies for identifying high-value AI assets, trust boundaries, and potential attack paths in complex AI environments
Capstone Red Team Engagement	Apply the techniques learned throughout the course during a full-spectrum red team engagement against a realistic enterprise AI environment, simulating how adversaries compromise production AI systems