AI Red-Teaming and AI Security Masterclass Training Course

Professional Development Training Course Outline

Category	Defense and Security	**Duration**	10 Days
Base Fee	$3,000 USD	**Delivery Mode**	Online / On-site Hybrid

Course Introduction

Artificial intelligence systems are transforming industries, yet they also introduce unprecedented security vulnerabilities that require proactive testing, ethical hacking, adversarial evaluation, and continuous risk mitigation. AI red-teaming has emerged as a critical discipline for identifying weaknesses in machine learning models, generative AI applications, and automated decision systems by simulating real-world threats and adversarial behaviors. AI Red-Teaming and AI Security Masterclass Training Course equips participants with cutting-edge knowledge in AI threat modeling, adversarial robustness, model manipulation, prompt-based attacks, and secure AI lifecycle management, ensuring organizations can build resilient and trustworthy AI systems.

As global adoption of AI accelerates, organizations must safeguard data pipelines, model outputs, and governance structures against misuse, bias exploitation, and malicious manipulation. This course provides hands-on techniques for identifying attack vectors, evaluating system vulnerabilities, and strengthening AI governance frameworks through practical red-team design, scenario testing, and advanced penetration methodologies. Participants will master the skills required to protect AI ecosystems, anticipate adversarial strategies, and implement security-by-design practices that uphold integrity, safety, transparency, and compliance in increasingly automated environments.

Programme Curriculum

AI Red-Teaming and AI Security Masterclass Training Course

Introduction

Artificial intelligence systems are transforming industries, yet they also introduce unprecedented security vulnerabilities that require proactive testing, ethical hacking, adversarial evaluation, and continuous risk mitigation. AI red-teaming has emerged as a critical discipline for identifying weaknesses in machine learning

models, generative AI applications, and automated decision systems by simulating real-world threats and adversarial behaviors. AI Red-Teaming and AI Security Masterclass Training Course equips participants with cutting-edge knowledge in AI threat modeling, adversarial robustness, model manipulation, prompt-based attacks, and secure AI lifecycle management, ensuring organizations can build resilient and trustworthy AI systems.

As global adoption of AI accelerates, organizations must safeguard data pipelines, model outputs, and governance structures against misuse, bias exploitation, and malicious manipulation. This course provides hands-on techniques for identifying attack vectors, evaluating system vulnerabilities, and strengthening AI governance frameworks through practical red-team design, scenario testing, and advanced penetration methodologies. Participants will master the skills required to protect AI ecosystems, anticipate adversarial strategies, and implement security-by-design practices that uphold integrity, safety, transparency, and compliance in increasingly automated environments.

Course Objectives

1. Understand core principles of AI red-teaming and adversarial testing.
2. Identify vulnerabilities across AI models, data pipelines, and deployment environments.
3. Apply trending AI security frameworks for governance and risk management.
4. Develop adversarial threat models aligned with real-world attack scenarios.
5. Execute prompt-based attacks on generative AI systems to assess robustness.
6. Analyze adversarial machine learning techniques and manipulation patterns.
7. Evaluate model bias, fairness risks, and exploit pathways using structured methods.
8. Apply monitoring, logging, and anomaly detection tools for AI systems.
9. Design secure model evaluation processes, stress-testing protocols, and validation methods.
10. Conduct red-team planning, execution, documentation, and reporting.
11. Integrate security-by-design into the entire AI lifecycle.
12. Strengthen organizational readiness for incident response and AI-related breaches.
13. Build long-term AI safety and security capacity through policy, culture, and training.

Organizational Benefits

- Enhanced protection of AI systems against adversarial threats
- Improved internal capacity for proactive AI risk detection and mitigation
- Strengthened governance and compliance with emerging AI regulations
- Increased resilience of machine learning models and datasets
- Reduced exposure to security breaches and model manipulation
- Better decision-making through robust AI assurance processes
- Increased trustworthiness and transparency of AI solutions
- Faster response to AI incidents, failures, and vulnerabilities
- Reduced operational losses caused by AI-driven risk events
- Improved competitive advantage through secure AI innovation

Target Audiences

- AI engineers and machine learning developers
- Cybersecurity and digital risk professionals
- Data scientists and AI researchers
- IT governance and compliance officers
- Digital transformation and technology managers

- Policy analysts and regulatory professionals
- Security auditors and penetration testers
- System architects and innovation leads

Course Duration: 10 days

Course Modules

Module 1: Foundations of AI Security and Red-Teaming

- Define AI red-teaming, its evolution, and relevance to modern security
- Understand categories of vulnerabilities across AI systems
- Explore global AI security standards and emerging regulations
- Assess risks in data pipelines, training sets, and model outputs
- Map typical AI threat actors, motives, and capabilities
- Case Study: Red-team discovery of hidden bias in a national AI deployment

Module 2: AI Threat Modeling and Risk Assessment

- Identify attack surfaces in ML and generative AI models
- Build threat models aligned with organizational risk exposure
- Evaluate adversarial capabilities and system weaknesses
- Apply structured techniques such as STRIDE and MITRE ATLAS
- Prioritize vulnerabilities for security interventions
- Case Study: Threat model for an automated digital lending AI

Module 3: Data Pipeline Security and Integrity Controls

- Assess vulnerabilities in data collection and preprocessing stages
- Implement validation, verification, and secure data handling
- Detect data poisoning, tampering, and injection risks
- Apply best practices for data provenance and auditability
- Strengthen controls for real-time and batch processing systems
- Case Study: Detecting data poisoning in a fraud detection model

Module 4: Adversarial Attacks on Machine Learning Models

- Understand gradient-based, black-box, and white-box attacks
- Analyze transferability and generalization of adversarial examples
- Test robustness of supervised and unsupervised models
- Apply evasion, extraction, and inference attacks
- Evaluate attack effectiveness and model degradation
- Case Study: Evasion attack on a credit scoring ML classifier

Module 5: Security in Generative AI and Prompt-Based Attacks

- Apply jailbreak and manipulative prompt techniques
- Test safeguards in chatbots, LLMs, and multimodal systems
- Identify prompt injection vulnerabilities
- Evaluate alignment, hallucination, and misuse risks
- Strengthen guardrails and monitoring mechanisms
- Case Study: Prompt injection exploit in a customer-facing LLM

Module 6: Model Robustness and Defensive Strategies

- Strengthen models through robustness enhancement techniques
- Apply regularization, adversarial training, and ensemble methods
- Conduct stress testing under multiple attack conditions
- Implement real-time detection of tampered inputs
- Evaluate defensive performance with red-team simulations
- Case Study: Adversarial training improving fraud model resilience

Module 7: Bias Exploitation, Fairness Risks, and Security Gaps

- Identify bias pathways vulnerable to exploitation
- Conduct algorithmic fairness assessments
- Map ethical risks associated with adversarial manipulation
- Apply quantitative fairness metrics and protections
- Red-team fairness vulnerabilities to strengthen governance
- Case Study: Bias exploitation in an AI-driven loan approval system

Module 8: AI System Architecture and Deployment Security

- Assess security of model hosting environments
- Secure APIs, endpoints, and integration interfaces
- Protect model artifacts and configuration files
- Implement network defenses for AI-based workflows
- Strengthen infrastructure using best practices
- Case Study: Architectural exposure in a cloud-hosted AI service

Module 9: Monitoring, Logging, and Anomaly Detection

- Set up continuous monitoring for AI security performance
- Build anomaly detection workflows for live AI systems
- Apply automated red flags for suspicious inference patterns
- Implement logging standards for traceability and transparency
- Use analytics dashboards for real-time security insights
- Case Study: Anomaly detection preventing AI system abuse

Module 10: Red-Team Exercise Planning and Execution

- Identify goals, scope, and methodologies for exercises
- Formulate red-team rules of engagement
- Conduct controlled adversarial experiments
- Document vulnerabilities, outcomes, and recommended fixes
- Present findings to leadership and governance bodies
- Case Study: Large-scale organizational AI red-team audit

Module 11: Incident Response for AI System Failures

- Detect and contain AI-related security events
- Apply escalation paths for model failures and adversarial attacks
- Document incident timelines and forensic evidence
- Implement recovery and rollback procedures
- Strengthen post-incident governance

- Case Study: Incident response to an AI misclassification crisis

Module 12: AI Governance, Compliance, and Accountability

- Examine emerging governance frameworks for AI security
- Integrate compliance into model lifecycle management
- Develop governance structures across teams and functions
- Apply transparency and documentation requirements
- Strengthen leadership oversight and accountability
- Case Study: Governance overhaul following AI audit findings

Module 13: Secure MLOps and Continuous Security Integration

- Integrate security protections into automated CI/CD pipelines
- Secure model deployment and retraining stages
- Implement policy checks and controls within MLOps tools
- Prevent unauthorized model updates and tampering
- Monitor pipelines for anomalies and vulnerabilities
- Case Study: MLOps security failure resulting in model drift

Module 14: AI Safety, Trust, and Responsible Use

- Evaluate safety risks in high-impact AI use cases
- Strengthen explainability, transparency, and human oversight
- Apply safety guidelines for socially sensitive AI applications
- Balance performance with risk mitigation
- Promote responsible adoption across the organization
- Case Study: Trust and safety breakdown in automated decision systems

Module 15: Scaling Secure AI Across the Enterprise

- Build long-term AI security maturity roadmaps
- Embed security-by-design culture across teams
- Evaluate enterprise-wide readiness through assessment tools
- Adopt best practices for secure AI procurement and vendor oversight
- Strengthen collaboration among technical and governance units
- Case Study: Enterprise transformation program for secure AI adoption

Training Methodology

- Instructor-led presentations supported by real-world AI security cases
- Hands-on exercises involving adversarial testing and prompt manipulation
- Scenario-based group work simulating red-team operations
- Model evaluation labs using attack and defense techniques
- Practical toolkits, templates, and risk assessment frameworks
- Peer-to-peer knowledge exchange and guided technical discussions

Register as a group from 3 participants for a Discount

Send us an email: info@fineskilltrainingcenter.org or call *+254769199797*

Certification

Upon successful completion of this training, participants will be issued with a globally- recognized certificate.

Tailor-Made Course

We also offer tailor-made courses based on your needs.

Key Notes

a. The participant must be conversant with English.

b. Upon completion of training the participant will be issued with an Authorized Training Certificate

c. Course duration is flexible and the contents can be modified to fit any number of days.

d. The course fee includes facilitation training materials, 2 coffee breaks, buffet lunch and A Certificate upon successful completion of Training.

e. One-year post-training support Consultation and Coaching provided after the course.

f. Payment should be done at least a week before commence of the training, to Fineskill Training Center account, as indicated in the invoice so as to enable us prepare better for you.

Upcoming Scheduled Sessions

Start Date	End Date	Location	Price
23 Nov 2026	04 Dec 2026	Online	$2,000
23 Nov 2026	04 Dec 2026	Physical	$3,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$6,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$9,000
23 Nov 2026	04 Dec 2026	Physical	$12,000
23 Nov 2026	04 Dec 2026	Physical	$12,000

Ready to enroll or request a customized program? Book online or contact us:

Register Online Now