KOENIG
step forward

AI Intelligence, Assurance and Defense in Cybersecurity

Duration (24h – 3 days)

DAY 1: AI Fundamentals & The Security Landscape: "Building the Foundation" | Duration: 8 hours

Session 1.1: Introduction to AI (90 minutes)
This session covers core AI concepts including definitions of Narrow AI vs. General AI, the relationship between Machine Learning, Deep Learning, and Generative AI, and key terminology such as models, training data, inference, and prompts. The focus is on building conceptual understanding without requiring programming knowledge.

Session 1.2: Understanding AI in Everyday Life (60 minutes)
Participants explore practical AI applications including enterprise chatbots, recommendation systems, fraud detection, and AI in GRC functions such as automated compliance monitoring and risk analytics. Software developers will learn about AI code assistants and automated testing tools.
Lab 1: AI Discovery Exercise (30 minutes)
Participants map AI tools currently used in their organizations using an interactive worksheet, creating an AI footprint map for their department.

Session 1.3: The AI Threat Landscape (90 minutes)
This critical session explains why AI systems are targeted, introduces the concept of attacks ON AI versus attacks WITH AI, and provides an overview of the MITRE ATLAS framework for AI threats. Real-world AI security incidents demonstrate business impact in terms GRC professionals and developers can relate to.

Session 1.4: Key Cybersecurity Concepts (60 minutes)
A security fundamentals refresher covers common attack vectors like phishing, malware, and social engineering, enterprise security controls, and how AI fundamentally changes traditional security models.

Session 1.5: The CIA Triad Applied to AI Systems (90 minutes)
The classic Confidentiality, Integrity, and Availability framework is recontextualized for AI systems: protecting training data (confidentiality), ensuring trustworthy outputs (integrity), and maintaining system performance (availability).
Lab 2: CIA Triad Mapping for AI Systems (Gap Analysis) (45 minutes)
Participants analyze an AI-powered customer service system, identifying risks to each CIA element and discussing appropriate controls through scenario-based worksheets.

DAY 2: AI in Cybersecurity - Threats, Attacks & Defenses: "Understanding the Battlefield" | Duration: 8 hours

Session 2.1: Hacking with AI - AI as a Weapon (60 minutes)

This session reveals how attackers leverage AI for reconnaissance, automated vulnerability scanning, AI-powered phishing, and the emerging threat of deepfakes. The goal is awareness of AI as an attack enabler.

Session 2.2: Penetration Testing with AI (60 minutes)

An overview of AI-assisted security testing tools covers ethical considerations and the GRC perspective on governing AI-powered security tools. Discussion focuses on when organizations should allow or restrict AI in security testing.

Session 2.3: Red Teaming with AI (60 minutes)

Participants learn AI red teaming concepts, including testing AI systems for vulnerabilities and an overview of the OWASP Top 10 for LLM Applications. Case studies from industry demonstrate real-world red teaming practices.

Lab 3: Collecting Red teaming/Vulnerability Assessment tools using AI. (60 minutes)

Session 2.4: Cryptography and AI (45 minutes)

This session covers AI's role in both breaking and strengthening encryption, post-quantum cryptography considerations, and future implications for data protection.

Session 2.5: AI in Cyber Defense - Security Operations (90 minutes)

AI-powered threat detection, SIEM integration, and Security Orchestration, Automation, and Response (SOAR) are explained conceptually. The focus is on balancing automation with human oversight.

Lab 4: AI Security Operations Simulation (60 minutes)

Using a web-based simulated SIEM/SOAR interface, participants analyze alerts from an AI-based threat detection system, prioritize incidents, and determine response actions while discussing when to trust AI recommendations versus human judgment.

Session 2.6: Cybersecurity Automation with AI (60 minutes)

Automation of repetitive security tasks, AI in vulnerability management, patch prioritization, and automated compliance checking are covered with practical examples.

Session 2.7: Threat Modeling for AI/ML Systems (60 minutes)

STRIDE methodology is applied to AI systems, covering trust boundaries across data, model, and inference layers. Participants complete a group activity identifying threats for a sample AI-powered HR recruitment tool.

DAY 3: AI Governance, Compliance & Ethics: "Building Responsible AI Practices" | Duration: 10 hours

Session 3.1: OWASP Top 10 for AI - An Overview (90 minutes)
Each of the OWASP LLM Top 10 vulnerabilities is explained: Prompt Injection, Insecure Output Handling, Training Data Poisoning, Model Denial of Service, and Supply Chain Vulnerabilities. Risks are mapped to business impact using language familiar to GRC professionals.

Lab 5: OWASP Risk Assessment Workshop (60 minutes)
Teams evaluate an AI-powered document classification system using the OWASP LLM Top 10 risk assessment checklist, producing a risk rating matrix with business impact scores and discussing prioritization based on risk appetite.

Session 3.2: Adversarial Attacks on AI Models (75 minutes)
Adversarial attacks are explained in simplified terms: evasion, poisoning, model extraction, and inference attacks. Real-world examples demonstrate image recognition failures and chatbot manipulation, with defense strategies including adversarial training, input validation, and continuous monitoring.

Session 3.3: Data Privacy in AI Systems (90 minutes)
GDPR requirements for AI systems processing personal data are covered, including data minimization, purpose limitation, the right to explanation, and automated decision-making provisions. Privacy by design principles for AI applications and India's DPDP Act considerations are also addressed.

Session 3.4: Ethical Frameworks and Guidelines for AI Development (90 minutes)
This session covers UNESCO's Recommendation on Ethics of AI, OECD AI Principles, and the EU AI Act risk classification. The FATE principles (Fairness, Accountability, Transparency, Explainability) are explained with case study discussions on ethical dilemmas in AI hiring, credit scoring, and content moderation.

Session 3.5: Ethical Considerations and Future of AI (75 minutes)
Responsible AI development practices include bias detection and mitigation strategies, human oversight models (Human-in-the-loop, Human-on-the-loop), and emerging regulations. Participants work on designing an AI governance policy outline for their organization.

Session 3.6: Prompt Engineering and Its Security Importance (75 minutes)
A non-technical introduction to prompt engineering explains why prompt design matters for security, how prompt injection attacks work, and defensive strategies.

Lab 6: Prompt Injection Awareness Lab (45 minutes)
Through a web-based safe sandbox, participants identify vulnerable vs. secure prompt designs, spot injections in sample prompts, and design basic system prompts with security guardrails. The focus is on recognition and defense concepts, not exploitation.

**timings mentioned for topics in the TOC are roughly estimated.