

**Understanding Generative AI Risks: Security, Ethics, and Social Implications**

**Course Number:** AI-148WA  
**Duration:** 1 day

**Overview**

As Generative AI (GenAI) becomes more widespread and impacts critical processes, understanding and managing its risks is crucial for successful operations. This risks of Generative AI training teaches attendees how to spot the dangers, risks, and vulnerabilities of GenAI. Participants learn to methodically examine GenAI systems for potential hazards and identify mitigations for those risks. In addition, attendees gain industry-leading resources to stay updated in this rapidly evolving area.

**Prerequisites**

This course is designed for personnel responsible for identifying, assessing, and managing the risks of Generative AI in their organization. It assumes they understand how Generative AI functions at a workflow level, including core steps in the training and prediction process.

**Materials**

All Generative AI Risks training students receive comprehensive courseware.

**Software Needed on Each Student PC**

Students should have Zoom installed as the conference platform.

**Objectives**

* Evaluate the ethical implications of AI systems, considering their potential impact on individuals, society, and the environment
* Understand the AI system lifecycle and the roles of different actors involved in its development, deployment, and use
* Apply the principles of AI ethics (safety, security, explainability, privacy, fairness, accountability, validity, and reliability) to the design and implementation of AI systems
* Identify and mitigate specific risks associated with generative AI, such as the creation of harmful content, privacy breaches, and bias
* Implement cybersecurity measures to protect AI systems from vulnerabilities and attacks, ensuring their security and resilience
* Recognize and address the potential negative consequences of AI failures, such as data breaches, misinformation, and discrimination
* Develop strategies to enhance the trustworthiness of AI systems, including explainability, transparency, and accountability
* Understand the emerging threats and vulnerabilities in AI cybersecurity and implement appropriate defenses
* Advocate for ethical considerations in AI research, development, and deployment to ensure a positive impact on society

**Outline**

* AI Ethics and Responsibility
  + What is an AI system?
  + The AI System Lifecycle
  + Common AI Actors
  + Principles of AI Ethics
  + Safe
  + Secure & Resilient
  + Explainable & Interpretable
  + Privacy-Enhanced
  + Fair
  + Accountable & Transparent
  + Valid & Reliable
  + Cybersecurity Triad & The Fallout of Failure
* GenAI Risks & Mitigations
  + Nefarious Information (e.g., CBRN)
  + Hallucinations
  + Dangerous or Violent Recommendations
  + Data Privacy
  + Environmental Impacts
  + Human-AI Configuration (e.g., workforce impact)
  + Information Integrity (e.g., misinformation)
  + Information Security
  + Intellectual Property
  + Obscene, Degrading, and Abusive Conduct
  + Toxicity
  + Bias
  + Homogenization
  + Supply Chain Integration
* GenAI Cybersecurity – Top 10 Vulnerabilities & Defenses
  + OWASP LLM Top 10
  + Prompt Injection
  + Insecure Output Handling
  + Training Data Poisoning
  + Denial of Service
  + Supply Chain Vulnerabilities
  + Sensitive Information Disclosure
  + Excessive Agency
  + Overreliance
  + Model Theft
* GenAI Cybersecurity – Tactics, Techniques, and Mitigations
  + MITRE ATLAS
  + Reconnaissance
  + Resource Development
  + Gaining Access
  + Execution
  + Persistence
  + Privilege Escalation
  + Defense Evasion
  + Credential Access
  + Discovery
  + Collection
  + ML Attack Staging
  + Exfiltration
  + Impact
* Conclusion
  + Frontier Threats
  + Additional Resources
  + Responsibility Matters