

**AI Security, Compliance, and Explainability**

**Course Number:** AI-114WA  
**Duration:** 2 days

**Overview**

This Artificial Intelligence (AI) Security, Compliance, and Explainability training course delves into the real-world applications and challenges shaping AI. Attendees learn the fundamentals of AI systems and core ethical principles, including fairness and transparency. Participants navigate global regulations, build secure AI models, and tackle bias at its root. This course gives your team the skills to shape a responsible, ethical, and secure AI future.

**Prerequisites**

Students should have foundational Knowledge in AI and Machine Learning, familiarity with Data Management, and understand basic Cybersecurity concepts.

**Materials**

All attendees recieve comprehensive couseware.

**Software Needed on Each Student PC**

Students should have Zoom installed as the conference platform.

**Objectives**

* Grasp the core principles of AI systems, their classifications, and their impact on various industries
* Understand the key ethical principles like fairness, accountability, and transparency, and learn to apply them in real-world AI scenarios
* Delve into the global regulatory landscape, from GDPR to specific industry regulations, and ensure your AI practices meet compliance standards
* Master AI cybersecurity, learn to mitigate data breaches and adversarial attacks, and build secure, trustworthy AI systems
* Understand Explainable AI (XAI) techniques and how AI models make decisions and build trust with transparent explanations
* Identify and address potential biases in AI systems, design fair and inclusive algorithms, and promote responsible AI development
* Collaborate effectively with AI systems, participate in AI auditing and certification processes, and contribute to building a responsible and beneficial AI future

**Outline**

* Introduction
* Ethics and Regulation
  + What is an AI System?
  + View of AI System
  + AI System Classifications
  + Branches of AI Today
  + AI by the numbers
  + AI - the Good
  + AI - the Bad
  + Principles of AI Ethics
  + Principles of AI Ethics
  + Fairness
  + Accountability
  + Transparency
  + Explainability
  + Privacy and autonomy
  + Reliable
  + Ask ChatGPT 3.5
  + AI Ethics in Practice
  + Regulatory Compliance in AI Systems
  + What are the benefits of AI regulation?
  + What are the disadvantages of regulating AI
  + Regulations and standards in AI
  + GDPR and data protection
  + AI in healthcare (HIPAA and other relevant laws)
  + AI in healthcare examples
  + AI in finance and regulatory compliance
  + US FINRA AI Deployment
  + AI in US finance examples
  + AI in the global finance examples
  + Case studies of AI non-compliance
  + Addressing Regulatory and Compliance
  + Dangers of Discrimination and Bias
  + Data Security and Data Privacy
  + Control and Security Concerns of AI
  + Cooperative Corporate Compliance
* Security and Privacy
  + What is AI Cybersecurity?
  + Threats and challenges in AI security
  + Implementing AI in cybersecurity
  + Adversarial attacks
  + Model inversion and extraction
  + Data poisoning
  + Best practices for securing AI systems
  + Robustness techniques
  + Differential privacy
  + Federated learning
  + Homomorphic encryption
* Secure AI Design and Deployment
  + Secure Software Development
  + Connectivity
  + Exploitation of AI Systems (Jailbreaks)
  + Infrastructure Concerns
  + System Vulnerabilities
  + Data Privacy
  + Data Leaks via Generating Text
  + OpenAI GPT-3/4 Data Location and Storage
  + Azure OpenAI
  + Adversarial Attacks
  + Malicious Use of AI
  + Bias and Discrimination
  + Regulatory and Ethical Considerations
  + Security and Privacy in Chatbots
  + Ensuring Security and Privacy
  + Data Protection
  + Enforcing Data Protection
  + Anonymization Techniques
  + Best Practices for Security with Generative AI
  + Sources of Bias in AI
  + Tackling AI Bias
  + Real-world Case Studies
  + Autonomous Vehicles and the Trolley Problem
  + AI in Warfare and Weaponization
  + AI in Criminal Justice
* AI Auditing and Certification
  + Introduction
  + Organizational Roles in AI Ethics and Compliance
  + Implementing AI Ethics Guidelines and Checklists
  + Key Components of an AI Audit
  + Steps in the AI Auditing Process
  + Post-Deployment Monitoring and Feedback Loops
  + Reporting and Recommendations
  + AI Certification Process
* Explainable AI (XAI)
  + Introduction to Machine Learning Interpretability
  + Importance of ML interpretability
  + Different types of ML interpretability models
  + Model-agnostic interpretability methods
  + Model-specific interpretability methods
  + Limitations of model-specific interpretability
  + Limitations of Model-agnostic interpretability
  + Global vs. Local interpretability
  + Interpretability in Deep Learning
  + Techniques and Methods for Explainability
  + Layer-wise relevance propagation (LRP)
  + Sensitivity analysis
  + Gradient-weighted class activation mapping (Grad-CAM)
  + Evaluating Interpretability
  + Techniques for evaluating interpretability
  + Overview of existing evaluation frameworks
  + Model-Agnostic Visual Analytics (MAVA)
  + Human-AI Collaborated Evaluation (HACE)
  + Interpretability in Large Language Models
  + Interpretability in Generative LLM’s
  + Common evaluation metrics for generative AI models
    - Diversity metrics
    - Likelihood
    - Perplexity
    - Inception Score
    - FID
    - BLEU
    - ROUGE
    - Human evaluation
  + Techniques for Interpreting Large Language Models
  + Importance of XAI in various sectors
  + XAI in Healthcare: Enhancing Care and Transparency
  + XAI in Finance: Driving Decisions and Building Trust
  + XAI in Legal Systems: Fairness and Accountability
* Conclusion