Module Title

Trustworthy AI

Module Type

In-person lecture course with supervised laboratory exercises

Credits

6 ECTS

Language of Instruction

English

Prerequisites

Basic knowledge of machine learning and deep learning, python programming, fundamental cybersecurity concepts is recommended.

Target Group:
This module is designed for Master's-level students in computer science with an interest in artificial intelligence, machine learning, or cybersecurity.

Module Description

This lecture provides a comprehensive introduction to the security, robustness, privacy, and safety of modern Artificial Intelligence (AI) systems, with a particular focus on deep learning. While AI systems are increasingly deployed in safety-critical and security-sensitive domains, they remain vulnerable to a broad spectrum of adversarial, privacy, and misuse-related threats. The course systematically studies these vulnerabilities and presents state-of-the-art defense mechanisms, risk management approaches, and protection techniques.

Students will learn how to analyze AI systems from a cybersecurity perspective across the entire AI lifecycle — from data collection and model training to deployment and post-deployment monitoring. The course integrates theoretical foundations with hands-on exercises, enabling students to implement attacks and defenses in practical machine learning settings.

Learning Objectives

After successful completion of the module, students will be able to:

- Explain core concepts of Artificial Intelligence, Deep Learning, and the AI lifecycle
- Identify and classify security, privacy, and safety risks in AI systems
- Understand threat models and adversarial capabilities in machine learning

- Implement and evaluate evasion, poisoning, and privacy attacks
- Analyze misuse and abuse risks in large language models (LLMs)
- Apply protection techniques such as adversarial defenses, watermarking, and secure aggregation
- Assess trade-offs between robustness, privacy, utility, and safety
- Critically evaluate emerging trends in trustworthy and secure AI

Course Content

The lecture is structured into eleven thematic units:

1. Foundations of AI and Deep Learning
Introduction to intelligence, artificial intelligence, deep learning architectures, training procedures, and the AI lifecycle.
Practical: Training a ResNet-18 model on CIFAR-10 using Google Colab.

2. Cybersecurity for Deep Learning
Security objectives, threat models, adversaries, and risk perspectives in AI. Introduction to the NIST AI Risk Management Framework and mapping security goals to lifecycle stages.

3. History and Taxonomy of Attacks on Deep Learning
Adversarial examples as a paradigm shift. Attack classification based on timing, goals, and capabilities (evasion, poisoning, privacy, abuse).

4. Evasion Attacks
In-depth analysis of adversarial examples, attack algorithms, and defenses.

5. Poisoning Attacks
Training-time attacks, data manipulation, backdoor insertion, and robustness strategies.

6. Privacy Attacks
Membership inference, model inversion, data extraction, and the privacy–utility trade-off.

7. Misuse, Abuse, and AI Safety
Jailbreaking, prompt injection, covert channels, dual-use concerns, and safety alignment in large language models.

8. Intellectual Property Protection & Model Ownership
Watermarking, fingerprinting, model extraction attacks, and legal/practical considerations.

9. Distributed Learning Security
Security challenges in federated learning, Byzantine robustness, secure aggregation, and privacy-preserving training.

10. Multi-Agentic AI and AI Assistants

This lecture examines the security, safety, and trustworthiness challenges of autonomous and semi-autonomous AI systems that interact with external tools, environments, and other agents.

11. Advanced and Emerging Topics
Certified robustness, supply chain attacks, backdoors in foundation models, red teaming, and regulatory developments.

Teaching and Learning Format

The course combines lectures with hands-on practical exercises. Students implement selected attacks and defenses in controlled experimental settings (e.g., CIFAR-10, MNIST, federated learning scenarios, and LLM environments). Multiple-choice assessments support conceptual understanding.
