AN TOÀN CHO AI (AI SECURITY)
Đề cương chi tiết (Course Syllabus)

1. General Information

Course name: An toàn cho AI (AI Security)

Course code: SEC1421_CLC

Course type: Selective

Number of credits: 2

2. Objectives

Knowledge:

The aim of this course is to provide students with fundamental and advanced knowledge of AI model attacks, AI privacy, and their corresponding defense strategies.

Skills:

On successful completion of this course a student will be able to:

- Analyze and identify security risks, vulnerabilities, privacy challenges and adversarial attacks targeting AI models
- Select and implement appropriate security measures to protect AI models, including privacy-preserving AI methods.

Attitude:

Students are required to attend the classes and complete assignments/projects.

3. Abstracts

This course provides students with basic and advanced knowledge of AI model attacks and their defense strategies, covering risks, vulnerabilities and privacy challenges in AI models, adversarial attack techniques, and solutions for prevention and response.

4. Teaching and learning methods

Lectures: 16h
Exercises: 4h
Projects: 6h
Labs: 4h
Individual reading: 0h

5. Prerequisites: Nhập môn trí tuệ nhân tạo – INT1341_CLC

6. Learning outcomes

After completing this course, the student is able to:
[LO1]: Explain AI security concepts, different types of AI model attacks, and the corresponding security measures to mitigate them.

[LO2]: Analyze and identify security risks, vulnerabilities, privacy challenges, and adversarial attacks targeting AI models.
[LO3]: Select and implement appropriate security measures to safeguard AI models against potential threats.

7. Assignment criteria

Learning outcomes	**Assignment criteria**
[LO1]: Explain AI security concepts, different types of AI model attacks, and the corresponding security measures to mitigate them.	Chapter 1, Chapter 2, Chapter 3
[LO2]: Analyze and identify security risks, vulnerabilities, privacy challenges and adversarial attacks targeting AI models.	Chapter 2
[LO3]: Select and implement appropriate security measures to safeguard AI models against potential threats.	Chapter 3

8. Outlines

Chapter 1: Introduction to AI Security
1.1. Security threats to AI
- 1.1.1. AI in critical systems
- 1.1.2. Security vs. reliability in AI models
- 1.1.3. The impact of AI security breaches

1.2. Categories of attacks on AI
- 1.2.1. Training-time attacks: Data poisoning, backdoors
- 1.2.2. Inference-time attacks: Adversarial examples, model extraction
- 1.2.3. Privacy threats: Membership inference, prompt injection, model inversion

1.3. Security challenges in Generative AI
- 1.3.1. Deepfake generation and detection
- 1.3.2. Model hallucinations and security implications
- 1.3.3. Privacy concerns in LLMs and AI-generated content

1.4. Examples of attacks and defenses in AI
- 1.4.1. Case study: Tesla autopilot adversarial attacks
- 1.4.2. Case study: AI-powered phishing & deepfake attacks
- 1.4.3. Case study: Membership inference attacks on medical AI models

Chapter 2: Adversarial Attacks on AI Models
2.1. Understanding Adversarial Examples
- 2.1.1. Definition and properties of adversarial perturbations
- 2.1.2. Targeted vs. untargeted attacks

2.2. Evasion Attacks
- 2.2.1. Fast Gradient Sign Method (FGSM)
- 2.2.2. Projected Gradient Descent (PGD)
- 2.2.3. Carlini-Wagner (CW) Attack

2.2.4. Adversarial attacks on LLMs
2.3. Data Poisoning Attacks (Training-time Attacks)
2.3.1. Label flipping and backdoor attacks
2.3.2. Data injection and poisoning techniques
2.4. Model Extraction & Model Inversion Attacks
2.4.1. Stealing AI models via API queries
2.4.2. Extracting sensitive information from AI outputs
2.4.3. Reconstructing training data from AI-generated outputs

Chapter 3: Defending AI Models
3.1. Adversarial Training
3.1.1. Training AI models with adversarial examples
3.1.2. Trade-offs between robustness and accuracy
3.2. Defensive Distillation
3.2.1. Reducing model sensitivity to adversarial noise
3.2.2. Implementation and limitations
3.3. Feature Squeezing & Input Preprocessing
3.3.1. Techniques to reduce adversarial noise impact
3.3.2. Practical applications and performance analysis
3.4. Detecting Adversarial Attacks
3.4.1. Adversarial detection using uncertainty estimation
3.4.2. Model confidence scoring
3.5. Defending against attacks on Generative AI
3.5.1. Detecting deepfake manipulations
3.5.2. Reducing LLM bias and manipulation risks
3.6. Privacy-Preserving AI Techniques and Governance
3.6.1. Privacy-Preserving AI Techniques
3.6.2. AI Privacy Governance & Legal Compliance

9. Required Textbooks

[1] Aneesh Sreevallabh Chivukula, Wan Lei Zhou, Bo Liu, Wei Liu, Xinghao Yang. Adversarial machine learning: attack surfaces, defense mechanisms, learning theories in artificial intelligence. Springer Nature, 2023.
[2] Ken Huang, Ben Goertzel, Jyoti Ponnapalli, Yale Li, Sean Wright, Wenge Yang. Generative AI Security: Theories and Practices. Springer, 2024.

10. Schedule

Main contents	Duration	Specific contents
Chapter 1: Introduction to AI Security	4h lecture 1h exercise	1.1. Security threats to AI 1.2. Categories of attacks on AI 1.3. Security challenges in Generative AI 1.4. Examples of attacks and defenses in AI

Chapter 2: Adversarial Attacks on AI Models	6h lecture 2h exercise 3h project 2h lab	2.1. Understanding Adversarial Examples 2.2. Evasion Attacks 2.3. Data Poisoning Attacks (Training-time Attacks) 2.4. Model Extraction & Model Inversion Attacks
Chapter 3: Defending AI Models	6h lecture 1h exercise 3h project 2h lab	3.1. Adversarial Training 3.2. Defensive Distillation 3.3. Feature Squeezing & Input Preprocessing 3.4. Detecting Adversarial Attacks 3.5. Defending against attacks on Generative AI 3.6. Privacy-Preserving AI Techniques and Governance

11. Grading Policy

Attendance:	10%
Mid-term exam/exercises:	10%
Course projects:	30%
Final examination:	50%