課程名稱(中文)：	人工智慧安全			開課單位：	資訊工程研究所
課程名稱(英文)：	AI Security			課程代碼：	4105109
授課教師：	阮文齡				
學分數：	3	必/選修：	選	開科年級：	碩博，大三年級四年級
先修科目或先備能力：					
課程概述：	Artificial Intelligence (AI) has been the key force in transforming our lives in the era of machine intelligence and automation. In the coming years, AI is expected to involve nearly every technology, e.g., Superintelligence (ChatGPT), AI for healthcare, manufacturing, autonomous vehicles and transportation systems, AI for agriculture, and environmental monitoring. However, AI creates new headaches for humans. For example, AI can be abused in disinformation campaigns or mishandled for harmful purposes, e.g., Deepfake, AI-empowered weapons, AI-empowered surveillance, and Cybercrime and hacking. This course will cover fundamental knowledge about AI security and attack/defense techniques on AI-empowered applications. Specifically, the introduction topics consist of (1) Basic applied AI/ML models; (2) Common threats/attacks in AI/ML (deep fake, adversarial attacks, data poisoning); (3) AI for threat hunting and attack defense; (4) AI tools for DevSecOps; (5) Security risks of superintelligence. Besides, AI techniques for solving some common tasks (e.g., checking bugs/security vulnerabilities, writing secure programs) are also introduced in this course. Finally, the principles of developing Responsible AI models to benefit humans are also discussed. Mid-term/final exam: Open book, but no electric device is allowed.				

學習目標：	● To allow students to acquire the basics of AI/ML models to use in common applications and be able to implement them in Python ● To allow students to acquire common threats and attack techniques against AI models ● To allow students to apply AI models/tools for threat hunting and attack defense ● To allow students to tweak several AI platforms for solving common security tasks ● To allow students to acquire the basics of Ethical AI in Cybersecurity
教科書：	No required textbook. Lecture slides are compiled by the teacher.

課程大綱		分配時數				核心能力	備註
單元主題	內容綱要	講授	示範	習作	其他		
Introduction	Course Overview & vision of AI security	3				☑1.1 ☑1.2 ☑1.3 ☐2.1 ☐2.2 ☐2.3 ☐3.1 ☐3.2 ☐3.3 ☐4.1 ☐4.2	
Applied AI	• Real-world AI applications	3				☑1.1 ☑1.2 ☑1.3 ☐2.1 ☐2.2 ☐2.3 ☐3.1 ☐3.2 ☐3.3 ☐4.1 ☐4.2	
Attacks against AI	• Data poisoning • Model inversion • Adversarial attacks • Deep fake & Disinformation attacks • AI-powered hacking	15				☑1.1 ☑1.2 ☑1.3 ☑2.1 ☐2.2 ☐2.3 ☑3.1 ☑3.2 ☑3.3 ☐4.1 ☐4.2	

AI defense	• Data sanitization • Model hardening • Adversarial training/Explainable AI • Anomaly Detection • AI-powered Threat Hunting	15	3			☑ 1.1 ☑1.2 ☑1.3 ☑ ☑2.1 ☐2.2 ☐2.3 ☐3.1 ☑3.2 ☑3.3 ☐4.1 ☐4.2	
Advanced topics	• Superintelligence and Prompt Engineering • Quantum Machine Learning • Project report	12	3			☑1.1 ☑1.2 ☑1.3 ☑2.1 ☐2.2 ☐2.3 ☑3.1 ☑3.2 ☑3.3 ☑4.1 ☑4.2	

1.1. 具有資訊工程相關基礎知識之吸收與了解的能力(Capability to grasp foundational knowledge in computer science.)

1.2. 具有運用資訊工程理論及應用知識，分析與解決相關問題的能力(Capability to use computer science theory and application knowledge to analyze and solve related problems.)

1.3. 在資訊工程的許多領域中，具有至少某一項專業能力，例如：硬體、軟體、多媒體、系統、網路、理論等。(Professional in at least one area, including hardware, software, multimedia, system, networking, and theory.)

2.1. 具有資訊工程實作技術及使用計算機輔助工具的能力。(Capability to perform computer science implementations and use computer-aided tools.)

2.2. 具有設計資訊系統、元件或製程的能力。(Capability to design computer systems, components, or processes.)

2.3. 具有科技寫作與簡報的能力。(Capability to write and present technical materials.)

3.1. 具有除了已有的應用領域之外，亦可以將自己的專業知識應用於新的領域或跨多重領域，進行研發或創新的能力。(Capability to apply one’ s professional knowledge to a new application domain or across multiple different application domains.)

3.2. 具有領導或參與一個團隊完成一項專案任務的能力並且具有溝通、協調與團隊合作的能力。(Capability to lead or participate in group projects, with effective communication, coordination, and teamwork.)

教學要點概述：
1. 教材編選：☑自編教材 ☐教科書作者提供
2. 教學方法：☑投影片講述 ☑板書講述
3. 評量方法： ☑上課點名 10%, ☐小考 0%, ☑作業 20%, ☐程式實作 0%, ☑ 實習報告 10% (bonus), ☐專案 0%, ☑期中考 40%, ☐期末考 0%, ☑期末報告 30%, ☐其它 0%
4. 教學資源：☑課程網站 ☐教材電子檔供下載 ☐實習網站
5. 教學相關配合事項：

課程目標與教育核心能力相關性	
請勾選：☐1.1 ☐1.2 ☐1.3 ☐2.1 ☐ 2.2 ☐ 2.3☐ 3.1 ☐ 3.2 ☐ 3.3☐ 4.1 ☐ 4.2	
1.1	**1.1 具有資訊工程相關基礎知識之吸收與了解的能力(Capability to grasp foundational knowledge in computer science.)**
	為何有關： AI is changing the way of solving many society and engineering challenges. The concepts and lab implementations covered in this course can help to explore AI capability and increase the awareness of students on AI risks and then create safe AI platforms.
	達成指標： Be able to understand and acquire the knowledge imparted in elective courses.
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points

1.2	**1.2 具有運用資訊工程理論及應用知識，分析與解決相關問題的能力 (Capability to use computer science theory and application knowledge to analyze and solve related problems.)**
	為何有關： Aside from taking advantage of CS abilities as traditional programming to create AI models, this course can let students leverage their own prompt programming skills and AI capability to solve coding and spoofing detection challenges.
	達成指標： Creative on creating AI models against security attacks and AI fake content generation
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points
1.3	**1.3 在資訊工程的許多領域中，具有至少某一項專業能力，例如：硬體、軟體、多媒體、系統、網路、理論等。(Professional in at least one area, including hardware, software, multimedia, system, networking, and theory.)**
	為何有關： To learn about fundamental AI security technologies and how to use prompt languages for work or detect false information.
	達成指標： Basic skill on evaluating the spoofing content/videos/materials
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points
2.1	**2.1 具有資訊工程實作技術及使用計算機輔助工具的能力。 (Capability to perform computer science implementations and use computer-aided tools.)**
	為何有關： To implement basic data poisoning and AI adversarial attacks as well as AI deep fake

	達成指標： Acquire the basic programming skill: using AI to generate AI code; Python programming to creating basic AI models.
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points
2.2	**2.2 具有設計資訊系統、元件或製程的能力。(Capability to design computer systems, components, or processes.)**
	為何有關：To create basic AI models for a variety of applications, such as traffic light/road lane detection, and then to attack them.
	達成指標：Basic skills on designing AI models for applications and corresponding attack strategies
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points
2.3	**2.3 具有科技寫作與簡報的能力。(Capability to write and present technical materials.)**
	為何有關：The writing and reading tasks will assist students in comprehending the complexities of an AI model and attack tactics in order to develop a defense mechanism.
	達成指標：Basic skills on writing report and findings from paper reading challenges.
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points
3.1	**3.1 具有除了已有的應用領域之外，亦可以將自己的專業知識應用於新的領域或跨多重領域，進行研發或創新的能力。(Capability to**

	apply one’s professional knowledge to a new application domain or across multiple different application domains.)
	為何有關：The course's hands-on projects and principles might help students launch new applications or propose effective defenses in other domains, such as computer vision.
	達成指標：Explore AI security knowledge for new applications or specific fields
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points
3.2	3.2. 具有領導或參與一個團隊完成一項專案任務的能力並且具有溝通、協調與團隊合作的能力。(Capability to lead or participate in group projects, with effective communication, coordination, and teamwork.)
	為何有關： The course tasks may necessitate teamwork and excellent communication with TA.
	達成指標： The habit to handle the assignments effectively and deliver them on time.
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points
3.3	3.3. 具有因應資訊科技快速變遷之能力，培養自我持續學習之能力。(Capability to adapt to rapidly changing computer science technology and to develop self-learning capabilities.)
	為何有關：Hands-on and reading challenges can assist students develop their self-learning abilities, particularly by keeping them up to date on the most recent attacks and defense tactics.
	達成指標：To learn the new attacks and defense techniques
	評量方法：

	Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points
4.1	**4.1. 具有社會責任、人文素養及奉獻精神。(The awareness of social responsibilities, humanity, and contribution.)**
	為何有關：The course's AI security topics will emphasize the necessity of responsible AI development and risk reduction.
	達成指標：To encourage students to contribute to Safe AI environments.
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points
4.2	**4.2. 具有工程倫理、宏觀能力、國際觀及前瞻視野。(The awareness of engineering ethics, broad capabilities, and global and contemporary vision.)**
	為何有關：Witnessing the risks of AI and their potential damages may motivate the students to use AI in an ethical manner and encourage the contributions to propose Trustworthy AI models.
	達成指標：To encourage students to utilize AI ethically
	評量方法： Level 5: Semester grade expected to be 90 and above Level 4: Semester grades can be expected to reach 80 points or above Level 3: Semester grades can be expected to reach 70 points or above Level 2: Semester grades can be expected to reach 60 points and above Level 1: Semester grades can be expected to be below 60 points