Subject Description Form

Subject Code	DSAI5210 / COMP5577
Subject Title	AI Security
Credit Value	3
Level	5
Pre-requisite/ Co-requisite/ Exclusion	Nil
Objectives	To equip students with a fundamental understanding of AI security and practical skills in evaluating and mitigating vulnerabilities in AI systems and leveraging AI tools and techniques to address common cybersecurity challenges. The objectives of this subject are to equip students to: 1. articulate the core concepts and principles underlying AI security; 2. identify and analyse various security threats to widely-used AI models; 3. comprehend and apply standard defense mechanisms to bolster the security of popular AI models; 4. gain proficiency in analysing and designing AI-based systems to tackle prevalent cybersecurity challenges; 5. cultivate practical skills for implementing attack and defense strategies to build secure AI systems that address typical cybersecurity issues.
Intended Learning Outcomes *(Note 1)*	Upon completion of the subject, students will be able to: <u>Professional/academic knowledge and skills</u> a) analyse and evaluate security threats associated with different phases of AI models in deployment; b) assess and execute various attacks against multiple AI models; c) apply and implement common defense mechanisms to enhance the security of diverse AI models; d) utilize AI techniques to address security-related challenges in common application domains. <u>Attributes for all-roundedness</u> e) acquire critical thinking and analytical skills, and improve technical writing as well as presentation skills.

<table>
<tr>
<td>Subject Synopsis/ Indicative Syllabus
(Note 2)</td>
<td>
<table>
<tr><th>Topics</th><th>Number of Lectures</th></tr>
<tr><td>1. Overview of AI Security
Introduce the fundamentals of popular AI models, principles of security analysis, security of AI, and AI for security.</td><td>2</td></tr>
<tr><td>2. Attacks on AI Model Training
Explore various attacks targeting the training phase of AI models, including data poisoning attacks and backdoor attacks, within domains such as computer vision, network analysis, and natural language processing.</td><td>2</td></tr>
<tr><td>3. Vulnerabilities of AI Models in Deployment
Analyse common threats encountered during the deployment of AI models, such as evasion attacks, membership inference attacks, model stealing attacks, and jailbreaking attacks, across fields like computer vision, network analysis, and natural language processing.</td><td>3</td></tr>
<tr><td>4. Common Defense Strategies
Examine widely-used defense techniques, including adversarial training, attack detection, data purification, and certified robustness, to strengthen the security of AI models at various stages of their lifecycle.</td><td>3</td></tr>
<tr><td>5. AI for Security
Design and analyse AI-based systems to address prevalent security challenges in various application domains, such as malware detection, network intrusion detection, and robust recommendation systems.</td><td>3</td></tr>
<tr><td>Total</td><td>13</td></tr>
</table>
</td>
</tr>
<tr>
<td>Teaching/Learning Methodology
(Note 3)</td>
<td>The course will be conducted through a blend of lectures, tutorials, student presentations, workshops, and a class project, emphasizing both the principles and practices of AI security. The foundational principles will primarily be addressed in lectures and tutorials, while practical aspects will be explored through workshops and projects. Given that AI security is an evolving field characterized by rapid advancements in research, students will be required to investigate selected topics and present their findings during lectures. The class project will provide an opportunity for students to reinforce their understanding, integrating both theoretical concepts and practical skills.</td>
</tr>
</table>

Assessment Methods in Alignment with Intended Learning Outcomes

(Note 4)

Specific assessment methods/tasks	% weighting	Intended subject learning outcomes to be assessed (Please tick as appropriate)				
		a	b	c	d	e
1. Assignments	15	√	√	√	√	
2. In-class Test(s)	25	√	√	√	√	
2. Class project	20	√	√	√	√	√
3. Examination	40	√	√	√	√	√
Total	100 %					

Explanation of the appropriateness of the assessment methods in assessing the intended learning outcomes:

Assignments: assessment of the studies with respect to the understanding of the relevant subject matters, including the principles, methodologies, and techniques by proving answers to the assignment questions.

Class project: assessment of the ability to solve real problems by using learned techniques and developing practical solutions.

In-class Test(s): assessment of the level of comprehension of core concepts and the ability to apply these concepts to analyse potential threats and solutions in AI systems.

Examination: assessment of the overall understanding of the course topics.

Student Study Effort Expected

Class contact:	
▪ Lecture	30 Hrs.
▪ Tutorial/Student Presentation/Workshop	9 Hrs.
Other student study effort:	
▪ Assignments + Class Project	35 Hrs.
▪ Self-study + Examination/Test Preparation	48 Hrs.
Total student study effort	122 Hrs.

Reading List and References

1. Vorobeychik, Y., & Kantarcioglu, M. (2018). *Adversarial machine learning*. Morgan & Claypool Publishers..

2. Miller, D. J., Xiang, Z., & Kesidis, G. (2023). *Adversarial Learning and Secure AI*. Cambridge University Press.

3. Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). *Deep learning* (Vol. 1, No. 2). Cambridge: MIT press.

4. Proceedings of IEEE Symposium on Security and Privacy

	5. Proceedings of USENIX Security Symposium 6. Proceedings of ISOC Network and Distributed System Security Symposium 7. Proceedings of ACM Conference on Computer and Communications Security 8. Proceedings of Machine Learning Research 9. Advances in Neural Information Processing Systems