CS5297: TOPICS IN AI SECURITY

New Syllabus Proposal

Effective Term
Semester A 2025/26

Part I Course Overview

Course Title
Topics in AI Security

Subject Code
CS - Computer Science
Course Number
5297

Academic Unit
Computer Science (CS)

College/School
College of Computing (CC)

Course Duration
One Semester

Credit Units
3

Level
P5, P6 - Postgraduate Degree

Medium of Instruction
English

Medium of Assessment
English

Prerequisites
Nil

Precursors
CS5486 Intelligent Systems or
CS5489 Machine Learning: Algorithms and Applications or
CS5491 Artificial Intelligence

Equivalent Courses
Nil

Exclusive Courses
Nil

Part II Course Details

Abstract

This course explores the intersection of artificial intelligence (AI) and cybersecurity, focusing on vulnerabilities, defense mechanisms, and ethical implications of AI systems. Students will explore adversarial attacks, data poisoning, privacy breaches, and model vulnerabilities while learning to design robust countermeasures like adversarial training, anomaly detection, and privacy-preserving techniques (e.g., federated learning). The course bridges technical challenges with ethical considerations, addressing bias, regulatory policies, and societal impacts.

Course Intended Learning Outcomes (CILOs)

	CILOs	Weighting (if app.)	DEC-A1	DEC-A2	DEC-A3
1	Analyze vulnerabilities in AI systems, identifying threats like adversarial attacks and data poisoning.	25	x	x	x
2	Design defense strategies to mitigate risks in AI models, including robust training and anomaly detection.	25	x	x	
3	Evaluate ethical and societal impacts of AI security, including privacy, bias, and regulatory compliance.	20	x	x	
4	Implement practical solutions using tools to secure AI systems in real-world case studies.	30	x	x	x

A1: Attitude
Develop an attitude of discovery/innovation/creativity, as demonstrated by students possessing a strong sense of curiosity, asking questions actively, challenging assumptions or engaging in inquiry together with teachers.

A2: Ability
Develop the ability/skill needed to discover/innovate/create, as demonstrated by students possessing critical thinking skills to assess ideas, acquiring research skills, synthesizing knowledge across disciplines or applying academic knowledge to real-life problems.

A3: Accomplishments
Demonstrate accomplishment of discovery/innovation/creativity through producing /constructing creative works/new artefacts, effective solutions to real-life problems or new processes.

Learning and Teaching Activities (LTAs)

	LTAs	Brief Description	CILO No.	Hours/week (if applicable)
1	Lectures	Students will explore the critical challenges of securing AI systems against emerging threats. Students will examine adversarial attacks, data poisoning, and privacy breaches while learning to design robust defenses using tools like TensorFlow Privacy and adversarial training frameworks. The course integrates technical rigor with ethical considerations, addressing bias, regulatory compliance, and societal impacts. Through case studies in healthcare, finance, and autonomous systems, and hands-on projects, students develop skills to protect AI models in real-world scenarios.	1, 2, 3, 4	2 hours/ week
2	Tutorials	Students will engage in hands-on, skill-focused sessions that reinforce weekly lecture content through practical application. Students engage in guided coding exercises (e.g., generating adversarial attacks, implementing defenses like adversarial training), analyze case studies (e.g., privacy breaches), and use industry tools such as PyTorch, TensorFlow Privacy, and the IBM Adversarial Robustness Toolbox. Activities emphasize real-world problem-solving, including securing AI models, evaluating robustness, and designing privacy-preserving systems.	1, 2, 3, 4	1 hour/ week

3	Project	Students will participate in collaborative group work and project, ensuring students bridge theory with practice. These projects directly align with course outcomes, equipping learners with technical expertise and critical thinking skills to address AI security challenges.	1, 2, 3, 4	2 hours/ week for 4 weeks

Assessment Tasks / Activities (ATs)

	ATs	CILO No.	Weighting (%)	Remarks ("-" for nil entry)	Allow Use of GenAI?
1	Assignment 1	1, 2	13	-	Yes
2	Assignment 2	2, 3	13	-	Yes
3	Assignment 3	3, 4	13	-	Yes
4	Project	1, 2, 3, 4	11	-	Yes

Continuous Assessment (%)

50

Examination (%)

50

Examination Duration (Hours)

2

Minimum Examination Passing Requirement (%)

30

Additional Information for ATs

For a student to pass the course, at least 30% of the maximum mark for the examination must be obtained.

Assessment Rubrics (AR)

Assessment Task

Problem set, including assignments and examination

Criterion

Ability to analyze fundamental AI security attacks and defences.

Excellent

(A+, A, A-) High

Good

(B+, B, B-) Significant

Fair

(C+, C, C-) Moderate

Marginal

(D) Basic

Failure
(F) Below Marginal

Assessment Task
Hands-on exercises

Criterion
Capacity to explore security toolkit and perform hands-on exercises, as well as explore the attack and defence technologies on software, system, and web.

Excellent
(A+, A, A-) High

Good
(B+, B, B-) Significant

Fair
(C+, C, C-) Moderate

Marginal
(D) Basic

Failure
(F) Below Marginal

Assessment Task
Project

Criterion
Ability to conduct a group project on selected security topics.

Excellent
(A+, A, A-) High

Good
(B+, B, B-) Significant

Fair
(C+, C, C-) Moderate

Marginal
(D) Basic

Failure
(F) Below Marginal

Part III Other Information

Keyword Syllabus

The syllabus will evolve over time as current topics change. Current topics will be selected from following. 1) Adversarial Attacks: Malicious inputs designed to deceive AI models (e.g., perturbed images fooling classifiers). 2) Data Poisoning: Corrupting training data to manipulate model behavior. Students learn detection methods (e.g., anomaly detection) and mitigation via data sanitization. 3) Privacy Preservation: Techniques like differential privacy and federated learning to protect sensitive data. 4) Robust Training: Hardening models against attacks by integrating adversarial examples into training. 5) Model Explainability: Ensuring transparency in AI decisions (e.g., SHAP values) to identify vulnerabilities and build trust. 6) Other topics in computer security: cloud security, security policy, information governance, information privacy, security evaluation, legal issues, computer crime and computer forensics, new access control paradigms, mobile Security, database security.

Reading List

Additional Readings

	Title
1	Ian Goodfellow et al., Deep Learning. (2017)
2	Nicolas Papernot et al., Security and Privacy in Machine Learning. (2018)