THE CHINESE UNIVERSITY OF HONG KONG
Print Course Catalog Details

Academic Org: Div of Computer Science & Engg – Subject: AI: Systems & Tech

Course: AIST5020 **Course ID:** 014552 **Eff Date:** 2024-07-01 **Crse Status: Active** **Apprv. Status:** Approved **[**New Course**]**
Trustworthy Artificial Intelligence 可信人工智能

This course introduces the principles and techniques of Trustworthy Artificial Intelligence (Trustworthy AI), which aims to mitigate the potential adverse effects of AI on people and society. The course focuses on four main aspects of trustworthy AI: privacy & security, robustness, explainability, and fairness. It covers the state-of-the-art research progress in these areas, including federated learning and adversarial attacks. Algorithms, models, and systems will be covered. Moreover, the course discusses the ethical and social implications of trustworthy AI, to foster social awareness among students who would use or develop AI techniques in the future. This course is suitable for students who have some background in machine learning, probability, and linear algebra.

這門課程介紹了可信人工智能（可信AI）的原則和技術，旨在減輕AI對人們和社會可能產生的不良影響。課程專注於可信AI的四個主要方面：隱私與安全、韌性、可解釋性和公平性。課程涵蓋了這些領域的最新研究進展，包括聯邦學習和對抗性攻擊。課程將涵蓋算法、模型和系統。此外，課程討論了可信AI的道德和社會影響，以培養學生對未來使用或開發AI技術的社會意識。該課程適合具有一定機器學習、概率和線性代數背景的學生。

Grade Descriptor:

A

EXCELLENT – exceptionally good performance and far exceeding expectation in all or most of the course learning outcomes; demonstration of superior understanding of the subject matter, the ability to analyze problems and apply extensive knowledge, and skillful use of concepts and materials to derive proper solutions.

有關等級說明的資料，請參閱英文版本。

B

GOOD – good performance in all course learning outcomes and exceeding expectation in some of them; demonstration of good understanding of the subject matter and the ability to use proper concepts and materials to solve most of the problems encountered.

有關等級說明的資料，請參閱英文版本。

C

FAIR – adequate performance and meeting expectation in all course learning outcomes; demonstration of adequate understanding of the subject matter and the ability to solve simple problems.

有關等級說明的資料，請參閱英文版本。

D

MARGINAL – performance barely meets the expectation in the essential course learning outcomes; demonstration of partial understanding of the subject matter and the ability to solve simple problems.

有關等級說明的資料，請參閱英文版本。

F

FAILURE – performance does not meet the expectation in the essential course learning outcomes; demonstration of serious deficiencies and the need to retake the course.

有關等級說明的資料，請參閱英文版本。

Equivalent Offering:
Units: 3 (Min) / 3 (Max) / 3 (Acad Progress)
Grading Basis: Graded
Repeat for Credit: N
Multiple Enroll: N
Course Attributes: MSc Computer Science
MPhil-PhD Computer Sci & Erg

Topics:

COURSE OUTCOMES

Learning Outcomes:

At the end of the course of studies, students will be able to:
1. Understand and explain the main concepts and techniques of trustworthy AI and its applications in various domains;
2. Compare and contrast the advantages and disadvantages of different trustworthy AI techniques;
3. Implement and evaluate different types of trustworthy AI techniques using popular frameworks such as PyTorch or TensorFlow;

4. Discuss the ethical and social implications of trustworthy AI, in terms of privacy, security, robustness, explainability, fairness, etc.;
5. Organize and conduct a group project for state-of-the-art research.

Course Syllabus:

Week 1. Introduction to trustworthy AI
Week 2. Privacy & Security: federated learning basics
Week 3. Privacy & Security: reconstruction attacks and inference attacks
Week 4. Privacy & Security: anonymization, noise injection, SMPC, and homomorphic encryption
Week 5. Robustness: outliers and adversarial attacks
Week 6. Robustness: adversarial detection and defense
Week 7. Explainability: ante-hoc explainability
Week 8. Explainability: post-hoc explainability
Week 9. Fairness: individual fairness and group fairness
Week 10. Fairness: fairness-aware federated learning
Week 11. Ethical and social implications of trustworthy AI
Week 12. Group presentation I
Week 13. Group presentation II

Assessment Type:

Homework or assignment	: 30%
Presentation	: 30%
Project	: 40%

Feedback for Evaluation:

1. Results of homework and assignments.
2. Course evaluation and questionnaire.
3. Reflection of teachers.
4. Question-and-answer sessions during class.
5. Student consultation during office hours or online.

Required Readings:

1. Trustworthy Machine Learning. Kush R. Varshney. Independently published. 2022.
2. A Survey of Trustworthy Federated Learning with Perspectives on Security, Robustness, and Privacy. Yifei Zhang, Dun Zeng,

Jinglong Luo, Zenglin Xu, Irwin King. ArXiv. 2023.

Recommended Readings:

1. Practical Adversarial Robustness in Deep Learning: Problems and Solutions. Pin-Yu Chen, Sayak Paul. CVPR 2021.

OFFERINGS

1. AIST5020 Acad Organization=CSEGV; Acad Career=RPG

COMPONENTS

LEC : Size=30; Final Exam=N; Contact=3
TUT : Size=30; Final Exam=N; Contact=1

ENROLMENT REQUIREMENTS

1. AIST5020

Enrollment Requirement Group:
For students in MSc Computer Science; or
For students in MPhil-PhD Computer Science & Engineering

New Enrollment Requirement(s):
Other Requirement = For students in MSc Computer Science; or
For students in MPhil-PhD Computer Science & Engineering

Additional Information

eLearning hrs for blended cls 0
VTL-Onsite face-to-face hrs 0
VTL-Online synch. hrs 0
VTL-Online asynch. hrs 0
No. of micro-modules 0