The Hong Kong University of Science and Technology (Guangzhou)

Title: AIAA5047 – Responsible AI

Credit(s): 3
Type: Elective
Prerequisite(s): N/A
Exclusion(s): N/A

Instructor: Sihong Xie
Email: sihongxie@hkust-gz.edu.cn
Office hours: *by appointment*

Lecture time: 9-11:50 AM Friday
Lecture location: Room 201, W2

Course Description

This course bridges the gap between technical and ethical concerns of modern AI, equipping students with a socio-technical mindset to design, critique, and govern AI systems that are responsible. Structured around five technical pillars—Foundational Models, Explainable AI (XAI), Uncertainty Quantification (UQ), Reinforcement Learning (RL), and Multi-Agent Systems (MAS)—each module explores real-world applications and trade-offs in high-stakes domains like healthcare, finance, robotics, etc. Students will learn to navigate challenges such as transparency, robustness, stakeholder alignment, and emergent system behavior, while recognizing that responsibility is not a constraint but the driving force of innovation. Graduates will emerge as leaders capable of building AI systems that inspire public trust, grounded in both technical rigor and a deep understanding of their societal impact.

Late work policy

You will have a total of **7 days** for late submission quota and you can use them freely for any HW and individual projects of your choice (any less-than-24-hours will be rounded up to a day). Due dates will be stated on all assignments. If we erroneously set conflicting dates across Canvas, the syllabus, and the assignment document, please inform us. Until any error is corrected the *earliest* date applies. Students are expected to be able to submit work correctly online and to back up their data. Therefore, “forgetting to click submit”, “computer crashes”, etc, are not acceptable lateness excuses. Note that online sites’ clocks may not match yours perfectly, so don’t wait until the last moment to submit.

Weekly Schedule

Week	Date	Topic
Week 1	2025/9/5	Course logistics Basic concepts of responsible AI
Week 2	2025/9/9	Introduction to foundational models (Transformer, GPT2, ViT, CLIP)
Week 3	2025/9/16	Transparency of AI: classic methods & LLM
Week 4	2025/9/23	Security, privacy, and bias of LLM
Week 5	2025/9/30	Uncertainty quantification basics
Week 6	2025/10/14	LLM uncertainty quantification
Week 7	2025/10/21	Visual model uncertainty quantification: 2D and 3D cases
Week 8	2025/10/28	Reinforcement learning: basics, exploration
Week 9	2025/11/04	Distributional RL; Sim2Real (tentative)
Week 10	2025/11/11	Safe planning using LLM and diffusion policy
Week 11	2025/11/18	Introduction to agentic systems with LLM
Week 12	2025/11/25	Safety of agentic systems
Week 13	2025/12/2	Final presentations

Assessment

Assessment Task	Contribution to Overall Course grade (%)	Due date
Homework	20%	See Canvas
Final exam	20%	
Final group project	30%	
Individual project	30%	

* Assessment marks for individual assessed tasks will be released within two weeks of the due date.

Grading Rubrics

Grading rubrics will be released after each homework assignment is graded. Students who have questions about their grades shall contact course Graduate Teaching Assistants (GTA) within ONE WEEK after the grades are released. After that, no appealing will be accepted.

Final Grade Descriptors

With the implementation of Outcome Based Education(OBE), the course adopts criterion-referenced assessment (CRA) and assign grades that reflect students' achievement. Specifically,

Grade	Score	Grade	Score
A+	>=95	B	>=75
A	>=90	B-	>=70
A-	>=85	C+	>=65
B+	>=80	C	>=60
		F	others

- https://www.hkust-gz.edu.cn/academics/academic-quality-manual/assessment/obe-ilos-and-criterion-referenced-assessment-cra/
- https://www.hkust-gz.edu.cn/academics/academic-quality-manual/assessment/grading-of-courses/

Course AI Policy

Students are encouraged to use AI tools to maximize the learning outcomes of this course.

Communication and Feedback

Assessment marks for individual assessed tasks will be communicated via Canvas within two weeks of submission. Students who have further questions about the feedback including marks should consult the instructor within five working days (one week) after the feedback is received.

Resubmission Policy

We do not allow resubmission of homework.

Required Texts and Materials

There is no required texts or materials. However, the instructor of each lecture may provide further optional texts or materials for students to learn more about the corresponding topic.

Academic Integrity

Students are expected to adhere to the university's academic integrity policy. Students are expected to uphold HKUST(GZ)'s Academic Honor Code and to maintain the highest standards of academic integrity. The University has zero tolerance of academic misconduct. Please refer to Regulations for Academic Integrity and Student Conduct for the University's definition of plagiarism and ways to avoid cheating and plagiarism.