香港浸會大學
HONG KONG BAPTIST UNIVERSITY

DEPARTMENT OF
COMPUTER SCIENCE
計算機科學系

DEPARTMENT OF COMPUTER SCIENCE

SEMINAR

2024 SERIES

Aligning Language Agents' Behaviours with Human Moral Norms

DATE & TIME

21 MAY 2024 (TUE) 10:00 – 11:00 AM

VENUE

Mr. and Mrs. Lee Siu Lun Lecture Theatre (WLB205),
The Wing Lung Bank Building for Business Studies, Shaw Campus

PROF. LING CHEN

Deputy Head of School (Research)
School of Computer Science
University of Technology Sydney

ABSTRACT

Language agents, designed to interact with their environment and achieve goals through natural language, traditionally rely on Reinforcement Learning (RL). The emergence of Large Language Models (LLMs) has expanded their capabilities, offering greater autonomy and adaptability. However, there's been little attention on augmenting the morality of these agents. RL agents are often programmed with a focus on specific goals, neglecting moral consequences, while LLMs might incorporate biases from their training data, which could lead to immoral behaviours in practical applications. This presentation introduces our latest research endeavours focused on enhancing both the task performance and ethical conduct of language agents involved in intricate interactive tasks.

For RL agents, we use text-based games as a simulation environment, mirroring real-world complexities with embedded moral dilemmas. Our objective thus extends beyond improving game performance to developing agents that exhibit moral behaviour. We first develop a novel algorithm that boosts the moral reasoning of RL agents using a moral-aware learning module, enabling adaptive learning of task execution and ethical behaviour. Considering the implicit nature of morality, we further integrate a cost-effective human-in-the-loop strategy to guide RL agents toward moral decision-making. This method significantly reduces the necessary human feedback, demonstrating that minimal human input can enhance task performance and diminish immoral behaviour.

Shifting focus to LLM agents, we begin with a comprehensive review of morality in LLM research, scrutinizing their moral task performance, alignment strategies for moral incorporation, and the evaluation metrics provided by existing datasets and benchmarks. We then explore how LLM agents can improve their moral decision-making through reflection. Our experiments, conducted within text-based games, show that integrating reflection enables LLM agents to make more ethical decisions when confronted with moral dilemmas.

SPEAKER'S BIOGRAPHY

REGISTER NOW

Enquiries: 3411-2385 **Email:** comp@comp.hkbu.edu.hk **Website:** https://bit.ly/bucs-events