THE HONG KONG POLYTECHNIC UNIVERSITY
香港理工大學

DEPARTMENT OF APPLIED MATHEMATICS
應 用 數 學 系

Seminar

Dr Xiaoyuan YI

Microsoft Research Asia

Topic
Value Compass: From Specific AI Risks to Basic Value Alignment

Date | Time
30th April 2024 (Tuesday) | 03:00 pm – 04:00 pm (HK Time)

Venue
Y301, PolyU

Abstract

Large Language Models (LLMs)' increasingly deep integration into human life bring potential risks to human society. To ensure the safe development of AI, it is necessary to regulate harmful content generated by AI models, such as toxicity, social bias, misinformation, etc. However, traditional methods of assessing and resolving risks become increasingly inadequate in the era of LLMs. Value Alignment becomes a more promising approach to fundamentally solve LLM risks. In this talk, we will first review the risks faced by LLMs, revisit the solutions designed for smaller models, and introduce the unique challenges they face in the era of LLMs. Subsequently, we will focus on one of the core technologies of LLMs, AI Alignment, introducing it from two perspectives: what to align and how to align, summarizing the challenges LLMs face in value alignment. We propose the Value Compass project, which is established from an interdisciplinary perspective, drawing on theories from ethics and social sciences, to handle problems related to definition, evaluation, and alignment of LLMs' values.

ALL ARE WELCOME