



“Math for Machine Learning” Seminar

Fundamental Limitation in Explaining AI (research presentation)

Prof. Atsushi Suzuki
HKU

Abstract

While large-scale models such as LLMs and diffusion models have achieved practical success, public institutions have emphasized the importance of explainability in AI. Existing methods for explaining AI, however, are not designed to provide completely faithful explanations of the behavior of large-scale AI systems. Although a completely faithful and interpretable explanation of the behavior of an AI system might be useful for AI governance, it has not been known whether providing such an explanation is theoretically possible. In this paper, with helps of information theory, we mathematically prove a fundamental quadrilemma in explaining AI, stating that AI and its explanation cannot satisfy the following four conditions simultaneously: 1) the complexity of the operation environment, 2) the goodness of the AI's performance, 3) the interpretability of the AI's explanation, and 4) the complete faithfulness of the AI's explanation.

Date:	September 10, 2026 (Thursday)
Time:	3:00 pm – 4:00 pm
Venue:	Room 210, Run Run Shaw Building HKU