Department of Psychology
The University of Hong Kong
香港大學心理學系

Brown Bag Lunchtime Seminar
(Theme: Cognition and Neuroscience)

Examining How Humans Learn from XAI Explanations with Implications for the Development of XAI with Theory of Mind Capacity

12:30 p.m. – 1:30 p.m. | May 12, 2023 (Friday)
Rm 813, 8/F, The Jockey Club Tower | Centennial Campus | The University of Hong Kong

Ruoxi QI
PhD student
Department of Psychology
The University of Hong Kong

Abstract

Recently, various eXplainable AI (XAI) methods have been developed to make the mechanisms of black-box AI models more transparent to users. However, most of these methods simply focused on using more AI to explain AI, without much consideration for the mental processes of the users. In order to develop XAI methods that can potentially acquire the same theory of mind (ToM) ability that human explainers have, it is necessary to examine how users process XAI's explanations and how they update their beliefs about the AI model accordingly. Therefore, the current study aims to investigate how users learn about the performance and strategy of an object detection AI model in driving scenarios based on saliency map explanations generated by XAI. The users' understanding of the AI model is measured by two tasks that assess simulatability, including the forward simulation task (predicting the AI's output) and the counterfactual simulation task (predicting the change to AI's output given a change to the input).

About the speaker

Ruoxi is a first year Ph.D. student supervised by Dr. Janet Hsiao. Her current research focuses on applying principles from psychology and cognitive science to the fields of AI and explainable AI.

Zoom (For participants who couldn't attend the Seminar in person)

https://hku.zoom.us/j/3951550048?pwd=SncvL3RYakEycUtpL29vdDJEdlEwdz09
Meeting ID: 395 155 0048 | Password: psyc

~All are Welcome~

Enquiry: rpsyc@hku.hk