NJIT

Department of Computer Science

Measuring artificial intelligence through a science of generalization: Perspectives from computer science, cognitive science, and neuroscience

Takuya Ito
IBM T.J. Watson Research Center

Hosted by: Cristian Borcea

Date: Friday, February 21, 2025
Coffee: 11:15 AM – 11:30 AM
Time: 11:30 AM – 12:30 PM (Eastern Time (US and Canada)
Location: GITC 4402 (4th floor Seminar Lecture Hall)

Zoom Link: **https://njit-edu.zoom.us/j/97433744280?pwd=FPIMXS9iE3xbFX8fa39baffc4fmlcQ.1**

Abstract:

The rapid development of modern artificial intelligence systems has created an urgent need for their scientific quantification. While current state-of-the-art models (e.g., LLMs) exhibit seemingly impressive cognitive capabilities, the sheer scale of model parameters and pretraining datasets make it difficult to discern the mechanisms that drive model capabilities. This makes it difficult, if not impossible, to assess whether current LLM behaviors are due to genuine cognitive capabilities or memorized routines. In this talk, I will describe research efforts aimed at characterizing the capabilities of AI generalization through computer science, cognitive science, and neuroscience perspectives. First, I will lay out a framework for measuring artificial intelligence by grounding generalization experiments inspired by compositionality and computational circuit complexity. Second, drawing on my prior research in cognitive science and neuroscience, I will describe recent research aimed at measuring human-addressable cognitive processes in AI systems via generalization. Third, I will describe past research characterizing the relationship between neural (biological and artificial) circuit representations and generalization. I will end with a discussion of how understanding the limits of model generalization facilitate the interpretable and appropriate usage of AI models in scientific applications, such as in neuroscience and psychiatry. This together outlines a research agenda for AI-driven science grounded through the lens of generalization.

Bio:

Takuya Ito is a Postdoctoral Researcher in the Mathematics & Theoretical Computer Science Department at the IBM T.J. Watson Research Center in Yorktown Heights, NY. At IBM, his research focuses on building frameworks to investigate reasoning, abstraction, and compositionality in artificial intelligence. Previously, he was a Swartz Fellow in Theoretical Neuroscience at Yale University, where he focused on understanding the neural mechanisms of cognitive flexibility in artificial and biological neural networks. He completed his PhD in Neuroscience at Rutgers University-Newark studying information representations in brain networks with human fMRI and monkey electrophysiology, and obtained his BA in Mathematics at Washington University in St. Louis with a minor in Computer Science. More information: https://ito-takuya.github.io/