Bocconi

6 OCTOBER 2023
15:00 PM

Bocconi University
Room 3-E4-SR03
Via Röntgen 1, Milano (3° floor)

“Controllability of Hallucinations in Large Language Models and AI Bots

Speaker

Stefano Soatto
Professor of Computing Science engineering
UCLA, Los Angeles

Abstract

Now that a significant fraction of human knowledge has been shared through the Internet, scraped and squashed into the weights of Large Language Models (LLMs), do we still need embodiment and sensory data to learn representations of the physical world? Can perceptual agents trained by passive observation learn world models suitable for decision and control? To begin tackling these questions, I will first address the issue of controllability of LLMs. LLMs are stochastic dynamical systems, for which the notion of controllability is well established: The state ("of mind") of an LLM can be trivially steered by a suitable choice of input given enough time and memory. However, the space of interest for control of an LLM is not that of words, but that of "meanings" expressible as sentences that a human could have spoken and would understand. Unfortunately, unlike controllability, the notions of meaning and understanding are not usually formalized in a way that is relatable to LLMs in use today. I will propose a simplistic definition of meaning that reflects the functional characteristics of a trained LLM. I will show that a well-trained LLM establishes a topology in the space of meanings, represented by equivalence classes of trajectories of underlying dynamical model (LLM). Then, I will describe both necessary and sufficient conditions for controllability in such a space of meanings.

Università
Bocconi
DEPARTMENT
OF COMPUTING
SCIENCES