B9145: Topics in Trustworthy AI

Course information and syllabus

Instructor: Hong Namkoong (namkoong@gsb.columbia.edu)
Lectures: Thursdays, 9am-12:15pm in Kravis 430
Office hours: By appointment
TA: Daksh Mittal (DMittal27@gsb.columbia.edu)

Description: Pre-trained AI systems have achieved remarkable capabilities in understanding videos, text, and code, demonstrating reasoning abilities that match or surpass human experts. While these omnipresent systems present unprecedented societal opportunities, significant challenges remain before they can meaningfully transform real-world decision-making problems.

A fundamental challenge is that AI systems inevitably encounter inputs unseen during training, as they must operate continuously while processing diverse real-world data including customer feedback and user interactions. Although scaling datasets has improved capabilities, it has not solved this core challenge. Modern AI systems, despite training on datasets orders of magnitude larger than human experience, still struggle with tail inputs – they hallucinate, cannot quantify uncertainty, and perform poorly on underrepresented groups.

The ability to handle tail inputs is a longstanding open problem in AI, with limited fundamental progress over past decades. As we exhaust easily available data sources, it is becoming clear that we must rethink the standard machine learning paradigm. Our lack of understanding of failure modes highlights the need for both more reliable models and rigorous safety evaluation methods.

This course surveys emerging topics in trustworthy machine learning, spanning data collection, pre-training, finetuning, and inference-time methods. Most topics discussed are active research areas, with reading materials drawn from recent literature (to be posted on the website). The goal is to foster discussion on new research questions, encompassing theoretical and methodological developments, modeling considerations, novel applications, and practical challenges.

Outline: The course will comprise of pedagogical lectures and seminar-style guided discussions. We will begin by overviewing recent advances in AI

I. Architectures, optimization algorithms, and datasets

II. Pre-training on web-scale data

III. Finetuning on downstream tasks, including supervised and RL-based methods

IV. Inference-time search methods

Then, we will cover the recent set of works on improving reliability in machine learning. Since trustworthiness is a loosely defined term with many connotations, we will explore various aspects of this concept, alongside a discussion of future directions. The following is a selection of topics that will be covered in the course (subject to change).

I. Data-centric view of AI systems

II. Distribution shift

III. Uncertainty quantification

IV. Adaptive data collection (active exploration)

V. Adversarial attacks

VI. Fairness, equity, and data provenance

VII. Causal learning

Prerequisites: There are no formal prerequisites, but the class will be fast-paced and will assume a strong background in machine learning, statistics, and optimization. This is a class intended for PhD students conducting research in related fields. Although some materials are of applied interest, this course has significant theoretical content that require mathematical maturity. The ability to read, write, and think rigorously is essential to understanding the material.

Grading and Evaluation:

- Final project (70%)
- Class presentation (30%)

Students taking the course for a grade will complete a final project for the course, which will count for 70% of the grade. Students are expected to work on an original research topic related to the content of the course, and at the end of the course the student(s) will present a brief writeup to the course staff detailing their work. Ideally, projects will have a chance to turn into publishable work.

In the case that progress on a research project prove difficult (and only when this turns out to be the case), students will have the option to do a pedagogical project. This can take the form of surveying the literature on a particular topic from a critical viewpoint, replicating the empirical work in a paper, or developing exercises from a few papers around topics in the class.

More information about the course project will be posted in the first few weeks of class. The project can be done individually, or in pairs. Students are expected to meet with the instructor during office hours to discuss their project ideas.