The Weizmann Institute of Science
Faculty of Mathematics and Computer Science

Vision and AI

Room 1, Ziskind Building
on Thursday, Dec 26, 2024
at 12:15

Niv Cohen
NYU

will speak on

Discovering and Erasing Undesired Concepts

Abstract:
The rapid growth of generative models allows an ever-increasing variety of capabilities. Yet, these models may also produce undesired content such as unsafe images, private information, or copyrighted material.

In this talk, I will discuss practical methods to prevent undesired generations. First, I will show how the challenge of avoiding undesired generations manifested itself in a simple Capture-the-Flag LLM setting, where even our top defense strategy was breached. Next, I will demonstrate a similar vulnerability in state-of-the-art concept erasure methods for Text-to-Image models. Finally, I will describe the notion of 'Unconditional Concept Erasure' aiming to mitigate such vulnerabilities. I will show that Task Vectors can achieve Unconditional Concept Erasure, and discuss the challenge of applying Task Vectors in practice.

Bio: Niv is a postdoctoral researcher at New York University hosted by Prof. Chinmay Hegde. He received a BSc in mathematics with physics as part of the Technion Excellence Program. He received his PhD in computer science from the Hebrew University of Jerusalem, advised by Prof. Yedid Hoshen. Niv was awarded the Israeli data science scholarship for outstanding postdoctoral fellows (VATAT). He is interested in anomaly detection, model personalization, and AI safety for Vision & Language models.