

Timothy H. Kostolansky

thkostolansky@gmail.com | tim0120.github.io

Education

Massachusetts Institute of Technology

Master of Engineering in Computer Science and Engineering

Bachelor of Science in Computer Science and Engineering

Bachelor of Science in Physics

Cambridge, MA

May 2024

May 2023

May 2023

Publications

CoT Red-Handed: Stress Testing Chain-of-Thought Monitoring

Benjamin Arnav*, Pablo Bernabeu-Pérez*, Nathan Helm-Burger*, Timothy H. Kostolansky*, Hannes Whittingham*, Mary Phuong

In Proc. NeurIPS 2025

Inverse Constitutional AI

Timothy Kostolansky

Master's thesis

Work and Research Experience

Center for Human-Compatible AI (CHAI), UC Berkeley

June 2025 – Present

Research intern, with Jiahai Feng

- Working towards an understanding of the geometry of entity representations in transformer language models
- Localizing and characterizing how language model activations represent entities and their properties
- Work in progress

London AI Safety Research (LASR) Labs

February 2025 – May 2025

Research fellow, with Mary Phuong (Google DeepMind)

- Tested Chain of Thought monitoring effectiveness under intentional subversion in an AI Control setup where models were instructed to pursue harmful side tasks while solving coding problems
- Evaluated CoT monitoring versus action-only monitoring for detecting adversarial model behavior, demonstrating that access to intermediate reasoning steps substantially improves detection of subtle deception
- Paper: <https://arxiv.org/abs/2505.23575>

Supervised Program for Alignment Research (SPAR)

June – October 2024

Researcher, with Jake Mendel (Apollo Research)

- Decomposed and reverse engineered neural networks that learn Boolean circuits
- Used linear probing, causal abstraction, and Boolean function measures (e.g., influence) in order to determine how small neural networks represent the parts of a Boolean circuit that it is trained on

MIT CSAIL, Algorithmic Alignment Group

July 2023 – September 2024

Graduate researcher for AI alignment lab led by Asst. Prof. Dylan Hadfield-Menell

- Developed and tested methods to extract a constitution which describes language models and preference datasets
- Used language modeling, clustering, textual semantic similarity, and contextual bandit methods to find a set of principles which describes a language model's behaviors in safety-relevant situations
- Thesis: <https://dspace.mit.edu/handle/1721.1/156804>

Second Spectrum Incorporated

June – August 2022

Software engineer for a sports data company that uses computer vision to track athletes in game film

- Upgraded and refactored video data pipelines from professional sports streams to the company's S3 servers
- Used Temporal.io to protect from failure over long-running protocols

Technical Skills

Languages: Python 3, Julia, MATLAB, Mathematica, TypeScript, bash

Tools: PyTorch, NumPy, pandas, ROS, React, Node.js

Extracurricular

MIT Science Policy Review

April 2021 – September 2023

Technology director for a policy journal that publishes science policy reviews authored by members of the MIT community

- Maintaining and updating the Review's website, uploading articles and covers

MIT Varsity Basketball

September 2019 – March 2022

NCAA Division III [athlete](#), competed with full course load, two-time NEWMAC Academic All-Conference selection

Japanese National Basketball Team

June 2019 – August 2019

- Selected for National Team and trained at Ajinomoto National Training Center in Tokyo
- [Competed](#) in the 2019 William Jones Cup in Taiwan, earned bronze medal

Non-Technical Skills

Languages: English (native), Japanese (proficient), Italian (learner)

Interests: meditation, chess, basketball, tennis, running, ortholinear keyboards