

Tim Hua

Website: timhua.me Alignment Forum: [Tim Hua](#)

EDUCATION

Middlebury College

BA, summa cum laude, Economics (GPA 3.97 / 4.00)

Middlebury, VT

May 2023

D.K. Smith Economics Prize for best thesis (*Fox News's Effect on Social and Moral Preferences*).

- Thesis cited by Harvard economist Benjamin Enke in [Moral Boundaries](#)

Harvard College

Visiting Undergraduate Student (GPA: 4.0 / 4.0)

Cambridge, MA

2021 - 2022 School Year

- Took coursework in economics, math, and statistics

RESEARCH

Steering Evaluation Aware Models to Act Like They Are Deployed [[arXiv](#), [Alignment Forum](#)]

Tim Hua, Andrew Qin*, Samuel Marks*, Neel Nanda**

NeurIPS 2025 Mechanistic Interpretability workshop

- Trained model organisms that would act differently in evaluation versus in deployment.
- Made AIs act like they are deployed using activation steering in settings where simple prompting fails.

Discovering Backdoor Triggers [[Alignment Forum](#)]

Andrew Qin, Tim Hua*, Samuel Marks*, Arthur Conmy*, Neel Nanda**

- Developed methods to uncover triggers to “semantic backdoors” in language models and had mixed results.

Combining Cost-Constrained Runtime Monitors for AI Safety [[arXiv](#), [Alignment Forum](#)]

Tim Hua, James Baskerville, Mia Hopman, Henri Lemoine*, Aryan Bhatt, Tyler Tracy*

NeurIPS 2025 (main conference)

- Developed theory and empirical demonstrations on how to optimally combine monitors with different cost and performance profiles.

AI Induced Psychosis: A Shallow Investigation [[LessWrong](#)]

Tim Hua

- Created simulated conversations between AIs and users experiencing psychosis and found concerned sycophantic behavior.
- Received media coverage from media outlets ([Business Insider](#), [Futurism](#)) and bloggers ([Zvi](#), [Tyler Cowen](#))

PROFESSIONAL EXPERIENCES

Neel Nanda and Sam Marks's MATS Scholar

Scholar

Berkeley, CA

Apr 2025 – Current

Tyler Tracy's MARS Mentee in AI Control

Researcher

Berkeley, CA

Jan 2025 – June 2025

Alice Rigg's AI Safety Camp Mentee in Mechanistic Interpretability

Researcher

Remote

Jan 2025 – June 2025

Walmart Economics Team

Manager, Economist

New York, NY

Intern in summer 2022, full time July 2023 – Aug 2024

Brookings Institution

Research Assistant for Professor Carol Graham

Remote

June 2021 – May 2022