Lovkush Agarwal

lovkush.com

AI and AI Safety Experience

Published Mech Interp paper in NeurIps Workshop. Arxiv pre-print — September 2024

- Studied how much information about the next paragraph is stored in '\n\n' token.

Project Lead for AI Safety Camp 10 — November 2024 – Apr 2025

- Projects asks how we can automate AI Safety research, e.g., can AI re-produce evals research in Inspect.

Teaching Fellow for BlueDot AI Alignment Course — October 2024 – Present

- Facilitating four cohorts. I am by far the most active facilitator on the Slack and have had more people join my cohorts. In particular, each week, I wrote an opinionated guide to the readings.

Adviser for CEEALAR — September 2024 – Present

- Review applications for their AI Safety Winter program.

Teaching Lead for ML4Good bootcamp — September 2024

- Responsible for the delivery of ML4Good, a 10-day in-person AI Safety bootcamp.
- Significantly improved the camp, e.g. my transformer lecture achieved highest student feedback, introduced a tutorial on Inspect, improved focus on meta-skills, creating a newsletter to improve the alumni community, etc.

Creating an LLM-playground app

- The user-friendly webapp helps non-technical people learn about LLMs, e.g. by trying SAE feature steering.

Writing about AI Safety

- E.g. Trying Anthropic's computer-use agent, Evals best practices and How I Keep Up With AI Safety.

Mentee for SPAR — June 2024 – September 2024

- Working with Nicky Pochinkov. Result was the NeurIps paper above.

Participant in AI Safety Hackathons

- Can we use Goodfire's API to recover their proprietary SAE feature vectors. Link to report.
- Running deception evals using AISI's Inspect open source framework. Code available on GitHub.

Judge for Impact Research Groups, Technical AI Safety

- See Impact Research Group's website for more information on their program.

Data Scientist (R&D), Shell — Apr 2021 – July 2024

- R&D for alignment of panel time series data. Includes lit reviews and designing new algorithms and metrics.
- Created a python package that allows geologists to align and measure uncertainty in well log correlations

Education

PhD in Pure Mathematics, University of Leeds, Link to thesis

MMath in Mathematics (Distinction), University of Cambridge, Link to thesis

Selected Position of Responsibility

President of The Cambridge University Mathematical Society

- Revived a stagnant society, by re-branding and obtaining sponsorship to fund weekly events for members
- Surpassed previous membership figures: increased from 50 per year to 200+ per year