Zhonghao He

☎ 44 07939466954 ✉ zh378@cam.ac.uk ◎ Cambridge
hezhonghao.github.io

SUMMARY

My research interests started with interpretability and AI alignment. On the one hand it’s about understanding the machines in front of us; on the other hand it’s about effective cooperation between humans and machines.

Ultimately, I want to build AIs for human excellence (or “arete”, in Greek conception), which requires both sound societal mechanism design and epistemic tools with which individuals can better exercise their agency.

I was awarded by Open Philanthropy's Graduate Scholarship. In a past life, I was an entrepreneur. I spent some time as COO of Charity Box and I started Homeal, which was funded by Hong Kong Government Cyberport.

EDUCATION

University of Cambridge Sep 2022 - Jul 2025

Mst in AI Ethics and Society Master Lucy Cavendish College Cambridge

- Coursework: Machine Learning Alignment, AI Ethics, AI Governance, History of AI, CS230 Deep Learning, Mathematics for Computer Science, ML Safety, Discrete Mathematics, CS234 Reinforcement Learning, Advanced Deep Learning Curriculum, CS109: Probability, Algorithm and Data Structure, Mechanistic Interpretability, Social Choice Theory, Game Theory, Category Theory.

Stanford Univerisity Jun 2019 - Sep 2019

Cognitive Science & Philosophy Summer Session Palo Alto

- Courses: Mathematics Foundation of Computing, Minds and Machines, Introduction to Neuroscience

Shantou University Aug 2014 - Jun 2019

English & Global Studies Bachelor Liberal Arts College Shantou

- Honors/Awards: Stanford Global Leadership & Engagement Program Scholarship, Hong Kong Cyberport Creative Micro Funding $100,000.
- Relevant Coursework: Machine Learning and relevant maths, Research Methodology, Linguistics.

RESEARCH EXPERIENCE

Cambridge & Berkeley Oct 2024 - Present

Co-lead

We started this new set of research on LLM's subtle influence on epistemic diversity and values of human society.

Cambridge & MIT

Project Lead

I lead a new research team investigating brain subjects (cognitive science & neuroscience) to address salient interpretability challenges (scalability, no benchmark, superposition, uninterpretable models).

Tentatively we aim for a Nature/Nature Machine Intelligence publication titled "What can ML interpretability researchers learn from neuroscience?"

See this ongoing work here.

Senior authors: Adrian Weller, Grace W. Lindsay

Peking University Aug 2023 - Nov 2023

David Krueger & Yaodong Yang's Labs — Beijing & Cambridge

I wrote an overview of interpretability for the purpose of safety and alignment (as part of a comprehensive alignment overview).

Senior authors: Yaodong Yang andSongchun Zhu

The paper is now on Arxiv, also this website: https://alignmentsurvey.com/

The framework (alignment circle) proposed in this paper was adopted by US's National Institute of Standards and Technology.

Cambridge AI Safety Labs — Dec 2022 - Present

Researcher — Cambridge

- Wrote a paper named "Harms from Increasingly Agentic Algorithmic Systems" .
- Accepted by ACM FAccT Conference; cited by GPT-4 technical report and high profile report such as "Managing AI Risks in an Era of Rapid Progress".
- As a major contributor, I participated in every stage of the paper from brainstorming to final editing. Specifically, I wrote/participated in section 2/3/4.

Center for AI Safety — Jun 2023 - Jul 2023

Research Assistant

- Contributed to "A REGULATORY FRAMEWORK FOR ADVANCED ARTIFICIAL INTELLIGENCE"

- Wrote one-pagers on a variety of AI topics for CAIS's policy work

Stanford University — Jun 2021 - Feb 2022

Research Fellow Stanford Exisential Risks Initiative

- Selected from over 300 applicants for a funded independent research project.
- My research focuses on the epistemic community framework and the global governance of AI. The article was later featured on Stanford's website.

Columbia University — Sep 2021 - Mar 2022

Research Assistant International Relations

- Working with Ph.D. candidate Jenny Xiao on a series of projects at the intersection of China, emerging technology (AI), and international cooperation. I used R to preprocess publication data of over 2000 samples, which were further used in the "Difference in Difference" analysis.

Concordia Consulting — Oct 2022 - Present

Affiliate Technical AI Safety Content & AI Governance Working Group — Remote

- Past projects: AI Alignment Review Chinese version; a submission to UN's Global Digital Compact; A clarification regarding FLI's open letter (3000 read on Chinese social media).
- Current projects: translation of a variety of technical AI safety work in the Chinese language.

Stanford Existential Risks Initiative (SERI) — Apr 2022 - Present

SERI Organizer — Beijing/Remote

- Built 1st program focusing on China's AI safety.
- Working with Open Philanthropy, and Longview Philanthropy, we successfully recruited 7 China-based, top STEM talents for AI safety research.
- 4 out of 7 fellows submitted an interoperability research paper to ICML.

Skills & Activities

- **Languages:** English (Fluent, TOEFL109, TEM 8 Certificate), Chinese (Native), French (Beginner)

- **Activities:** Cambridge Union, Cambridge Technology Society, Lucy Cavendish Boat Club, Cambridge China Forum.
- **Interests:** Debate, Bodybuilding, Hiking, Rowing, Effective Altruism, Greeks, Nietzsche.
- **Skills:** Python (fluent), Pytorch, ML/AI, Interpretability techniques, Matlab, Web stuff, Data Visualization, R.