

# ZHONGHAO HE

zh378@cam.ac.uk | hezhonghao.github.io | github.com/hezhonghao | [Google Scholar \(500+citations\)](#)

## SUMMARY

---

I am Zhonghao. I work on AI alignment and human-AI interaction research. My previous work got accepted by NeurIPS, ICML, ACM FAccT, and ICLR (workshop), etc. My major interests are to build machines that help humans learn and think. Currently I focus on two things, to develop truth-seeking AI (Bayesian & coherent & making discoveries), and to solve “positive feedback loop” problems in tech products: LLM sycophancy, confirmation bias in reasoning models, social media echo chamber, and polarization. [tinyurl.com/prevailai](https://tinyurl.com/prevailai)

## RESEARCH EXPERIENCE

---

**Cosmos Fellow / Research Associate - Oxford** *Jan 2026 - Dec 2026*

Admitted as Cosmos Fellow of HAI Lab, Oxford. Working with Prof Philipp Koralus of Philosophy Department and Prof Jakob Foerster of Engineering Department. Projects: Martingale training for Bayesian rationality, Coherence optimization for self-improvement, Coherence game for learning human reflective equilibrium, and infrastructure for truth-seeking AI.

**Research Engineer - CMU (Remote)**

*Jan 2025 - Dec 2026*

Co-lead “Martingale Score”: We introduce a Bayesian statistical method to evaluate confirmation bias in LLM reasoning, with Profs Maarten Sap & Hirokazu Shirado [Link to Paper](#)

**Research Engineer - University of Washington (Remote)**

*Oct 2024 - Jun 2025*

Co-led two papers: “The Lock-in Hypothesis”, and “Open Problems in AI Influence”, with Prof Max Kleiman-Weiner [The Lock-in Hypothesis Website](#)

**Researcher - University of Cambridge**

*Dec 2023 - Jul 2025*

Worked on multiple projects on interpretability, alignment, and agentic safety, with Profs David Krueger, Yaodong Yang, Grace W. Lindsay, and Anya Ivanova.

## EDUCATION

---

**University of Cambridge**

*Sep 2022 - Jul 2025*

*Mst in AI Ethics*

*Coursework: ML Safety, AI Alignment, AI Ethics, RL, Advanced DL, Algorithm and Data Structure, Mechanistic Interpretability, etc.*

**Stanford University**

*May 2019 - Aug 2019*

*Cognitive Science Summer Semester*

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

**Shantou University**

*Aug 2014 - Jun 2019*

*BA in English and Linguistics*

*Relevant Coursework: Linguistics, ML, Maths.*

## AWARDS AND GRANTS

---

**Cosmos Fellow**

*2026*

**UK AISI Alignment Project Finalist (Recommended by AISI to funding partners)**

*2025*

**Cosmos Grant on Truth-seeking AI**

*2025*

**Foresight Institute AI Safety Research Grant**

*2025*

## PUBLICATIONS

---

- [1] **Z. He\***, T. Qiu\*, H. Shirado, M. Sap (2025) Stay True to the Evidence: Measuring Belief Entrenchment in LLM Reasoning via the Martingale Score. *NeurIPS 2025*.
- [2] T. Qiu\*, **Z. He\***, T. Chugh, M. Kleiman-Weiner (2025). The Lock-in Hypothesis: Stagnation by Algorithm. *ICML 2025*.
- [3] **Z. He\***, T. Qiu\*, T. Lin, M. Glickman, J. Wihbey, M. Kleiman-Weiner (2025). Position: AI Systematically Rewires the Flow of Ideas. *ICLR 2025 BiAlign Workshop*.
- [4] **Z. He\***, M. Tehenan\*, J. Achterberg, K. Collins, K. Nejad, D. Akarca, Y. Yang, W. Gurnee, I. Sucholutsky, Y. Tang, R. Ianov, G. Ogden, C. Li, K. Sandbrink, S. Casper, A. Ivanova, G. W. Lindsay (2024). Multilevel interpretability of artificial neural networks: leveraging framework and methods from neuroscience.
- [5] J. Ji, T. Qiu, B. Chen, B. Zhang, H. Lou, K. Wang, Y. Duan, **Z. He**, J. Zhou, Z. Zhang, F. Zeng, K. Y. Ng, J. Dai, X. Pan, A. O’Gara, Y. Lei, H. Xu, B. Tse, J. Fu, S. McAleer, Y. Yang, Y. Wang, S. C. Zhu, Y. Guo, W. Gao (2023). AI Alignment: A Comprehensive Survey. Under review at ACM Computing Surveys.
- [6] A. Chan, R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krashenninnikov, L. Langosco, **Z. He**, Y. Duan, M. Carroll, M. Lin, A. Mayhew, K. Collins, M. Molamohammadi, J. Burden, W. Zhao, S. Rismani, K. Voudouris, U. Bhatt, A. Weller, D. Krueger, T. Maharaj (2023). Harms from increasingly agentic algorithmic systems. *Accepted by ACM FAccT 2023*

## PROFESSIONAL SERVICES

---

### Invited Talks:

- Nov 2025 META FAIR
- Nov 2025 MIT
- Oct 2025 UK AI Security Institute
- Oct 2025 University of Chicago
- Sep 2025 Tsinghua University
- Jul 2025 University of Washington
- Feb 2025 Cambridge University

### Mentoring:

- Jul 2025 – Oct 2025 Supervised Program for Alignment Research
- Jul 2025 – Oct 2025 Algorverse AI Safety Fellowship

### Reviewing

Nov 2025 - IASEAI 2026

2025 Onwards - Transactions on Machine Learning Research (TMLR)