ZHONGHAO HE

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

SUMMARY

I am Zhonghao, a master's student at the University of Cambridge. I work on AI alignment, interpretability, and human-AI interaction research. My previous work got accepted by ICML, ACM FAccT, and ICLR (workshop). My major interests are to design machines that help humans learn and think. Currently I focus on two things, to develop truth-seeking AI (Bayesian & exploring truth), and to solve "positive feedback loop" problems in tech products: LLM sycophancy, confirmation bias in reasoning models, social media echo chamber, and polarization. https://tinyurl.com/projectprevail

RESEARCH EXPERIENCE

Advisor - Algoverse (Remote) *July 2025 - Present*
Co-mentor of two research projects on truth-seeking AI, Martingale training, and measurement of LLM's societal impact.

Research Engineer - MIT (Remote) *Jun 2025 - Present*
Co-lead a project to use RL to infer human goals and levels of comprehension, with Prof Andreea Bobu.

Research Engineer - CMU (Remote) *Jan 2025 - Present*
Co-lead "Martingale Score": We introduce a Bayesian statistical method to evaluate confirmation bias in LLM reasoning, with Profs Maarten Sap & Hirokazu Shirado A manuscript.

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 - Present*
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 Grant on Truth-seeking AI *2025*
Foresight Institute AI Safety Research Grant *2025*
Lambda Research Grant *2024*
Manifund Research Scholarship *2023*
Open Philanthropy's Graduate Scholarship *2022*

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. *Under Review.*

[2] T. Qiu*, **Z. He***, T. Chugh, M. Kleiman-Weiner (2025). The Lock-in Hypothesis: Stagnation by Algorithm. *Accepted to 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. Krasheninnikov, 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*