Updated October 2026

Muhua (黄牧华) Huang

muhua@stanford.edu · muhua-h.github.io

I am an ex-student researcher at **Google DeepMind** and **Microsoft Research**. I study how individuals come together to form shared **values**, **cognition**, and **culture**, and what changes when **AI systems** join this process. My work asks how AI differs from **human intelligence**; what large language models, as compressed **collective human knowledge**, reveal about **human behavior and society**; and how AI, as a **social actor**, reshapes **social dynamics** and **power structures**.

Research: Computational Social Science, Organizational Behavior, AI Alignment, Human-AI Collaboration
Engineering: GABM, Large-Scale LLM Experimentation, Fine-tuning, Agents, High-Performance Computing
Analytical: Psychometrics, LLM Evaluation, Embeddings, Network Analysis, ML/NLP, Behavioral Experiments

Education

Stanford Graduate School of Business 2025 – 2030
PhD in Organizational Behavior

The University of Chicago 2023 – 2025
MA in Computational Social Science

The University of British Columbia (UBC) 2018 – 2023
BA in Computer Science & Psychology (Honours)

International AI Evaluation Programme Fellow 2026

Summer Institute in Computational Social Science (SICSS) 2024

Google Research Mentorship Program Scholar 2020 – 2021

Publications

[1] **Huang, M.**, & Evans, J. (2026). Institutions as cached computation for resource-rational negotiation. *Behavioral and Brain Sciences*, 49, e144. DOI

[2] **Huang, M.**, Zhang, X., Soto, C., & Evans, J. (2026). Designing AI-Agents with Personalities: A Psychometric Approach. *Personality Science*. DOI

[3] Bai, Y., Duan, S., **Huang, M.**, Yao, J., Liu, Z., Zhang, P., …& Xie, X. (2026). IROTE: Human-like Traits Elicitation of Large Language Model via In-Context Self-Reflective Optimization. *Proceedings of the AAAI Conference on Artificial Intelligence*. DOI

[4] Savalei, V., & **Huang, M.** (2025). Fit Indices Are Insensitive to Multiple Minor Violations of Perfect Simple Structure in Confirmatory Factor Analysis. *Psychological Methods*. DOI

[5] Zhang, X., **Huang, M.**, Sun, J., & Savalei, V. (2025). Improving the Measurement of the Big Five via Alternative Formats for the BFI-2. *Journal of Personality Assessment*. DOI

[6] Yao, J., Yi, X., Duan, S., Wang, J., Bai, Y., **Huang, M.**, …& Xie, X. (2025). Value compass benchmarks: A comprehensive, generative and self-evolving platform for LLMs' value evaluation. *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*. DOI

[7] Kim, H., Yi, X., Bak, J., Yao, J., Lian, J., **Huang, M.**, Duan, S., & Xie, X. (2025). The Road to Artificial SuperIntelligence: A Comprehensive Survey of Superalignment. *SuperIntelligence - Robotics - Safety & Alignment*, 2(1). DOI

[8] Laurin, K., Engstrom, H., & **Huang, M.** (2024). What will my life be like when I'm 25? How social class predicts kids' answers to this question, and how their answers predict their futures. *Journal of Social Issues*. DOI

Working Papers & Manuscripts in Preparation

[1] **Huang, M.**, & Guilbeault, D. (Manuscript in preparation). AI as Alien Intelligence: Shared Vocabulary Masks Divergent Category Systems in Human and AI Collectives. *Manuscript available upon request.*

[2] How Cultural Narratives about AI Shape AI Behavior and Alignment: Measuring and Removing Narrative-Installed Personas in Language Models. Manuscript in preparation.

[3] Lattice: Measuring Horizontal and Transferable Skills in AI Agents. Google DeepMind. Manuscript in preparation.

[4] Zhang, H., **Huang, M.**, & Wang, J. (Under review). Computational Multi-Agents Society Experiments: Social Modeling Framework Based on Generative Agents. [Preprint]

Conference Presentations

† first author / * solo

- **Designing AI-Agents with Personalities: A Psychometric Approach**
  [*]APA, Seattle (Aug 2024, Poster); *CPA*, St. John's (Jun 2025, w/ X. Zhang); *CPA*, Montréal (Jun 2026, w/ X. Zhang); *ISDSA*, Beijing (Jul 2026, Speed Talk, w/ X. Zhang)
- **Beyond Anthropomorphism: Unveiling Unique Value Structure of Large Language Models** (w/ P. Biedma, X. Yi, L. Huang, M. Sun, J. Evans, X. Xie)
  [†]IC2S2, Norrköping (Jul 2025)
- **Re-Discovering the Big Five: Using LLM Embeddings to Understand Personality Structure** (w/ N. Huang)
  [†]APA, Denver (Aug 2025)
- **Improving the Measurement of the Big Five via Alternative Formats for the BFI-2** (w/ X. Zhang, V. Savalei)
  [†]APA, Washington DC (Aug 2023, Poster); *IMPS*, Minneapolis (Jul 2025, w/ J. Sun); *CPA*, Montréal (Jun 2026)
- **Social Class Differences in Children's Hopes and Expectations for the Future** (w/ H. Engstrom, K. Laurin)
  [†]APS (May 2021, Poster); *SPSP*, San Francisco (Feb 2022, Talk); *Society in the Classroom*, London (Jul 2022, Talk)
- **Understanding Silicon Sampling through a Psychometric Lens** (w/ C. F. Falk, X. Zhang, N. Guenole, W. Li)
  CPA, Montréal (Jun 2026)
- **Can SEM Fit Indices Distinguish Between CFA and EFA Data Structures?** (w/ V. Savalei)
  SMEP, Iowa City (Oct 2023)
- **Gauging Student Engagement with an XAI Interface via Eye-tracking** (w/ C. Conati, R. Murali)
  IJCAI Workshop on XAI (Jun 2022)

Invited Talks

[1] Yang, L., Sun, L., **Huang, M.**, Wang, J., Jiang, R., & Xu, F. (November 14, 2024). Panel discussion: LLM-driven social science and generative agents. *2024 MSR Asia TAB Workshop: Shaping the Future with Societal AI*. Microsoft Research Asia, Beijing, China.

[2] **Huang, M.** (January 13, 2025). Designing LLM-agents with personalities: A psychometric approach. Invited talk at the *Quantitative Methods Forum*, York University, Toronto, Canada.

Industry Research Experience

Student Researcher | Google DeepMind 2026.06 – 2026.09
Advisor: Martin J. Gonzalez

Contributing Researcher and Writer | Stanford HAI & Google DeepMind 2026.01 – 2026.04
Advisor: Martin J. Gonzalez

Research Intern | Microsoft Research Asia 2024.07 – 2024.10
Advisors: Dr. Xiaoyuan Yi and Dr. Xing Xie

Academic Research Experience

PhD Student Researcher | Stanford Graduate School of Business 2025.10 – Present
Advisor: Dr. Douglas Guilbeault

Master's Thesis | University of Chicago 2023.09 – 2025.06
Advisor: Dr. James Evans

Honours Thesis, Quantitative Psychology | University of British Columbia 2020.04 – 2023.08
Advisor: Dr. Victoria Savalei

Honours Thesis, Social Psychology | University of British Columbia 2020.08 – 2021.09
Advisor: Dr. Kristin Laurin

Research Assistant, Human-Centered AI | University of British Columbia 2021.10 – 2022.09
Advisor: Dr. Christina Conati

Teaching & Service

Computational Social Science Workshop Coordinator | University of Chicago 2024.10 – 2025.06

Deep Learning Textbook Editor | University of Chicago 2023.10 – 2024.10

CS Teaching Assistant | University of British Columbia 2021.01 – 2022.12

Crisis Respondent | Kids Help Phone 2019.09 – 2020.08

Peer Reviewer | *PNAS Nexus*; *Scientific Reports*; *Personality Science*; *Behaviour & Information Technology*

Honors & Awards

Stanford Graduate School of Business Doctoral Fellowship	2025 – 2030
Stanford University EDGE Doctoral Fellowship	2025
UChicago Outstanding Thesis Award	2025
Microsoft Research Asia Outstanding Intern Award (Stars of Tomorrow Certificate)	2024
OpenAI Researcher Access Program	2024
SICSS-Beijing Merit-Based Scholarship	2024
UChicago Computational Social Science Research Poster Competition (Honorable Mention, 2nd Place)	2024
UChicago Quadrangle Research Scholarship	2023, 2024
UChicago Social Sciences Promise & Merit Scholarship	2023, 2024
Quinn Research Assistantship Award	2021
UBC International Community Achievement Award	2021
UBC Trek Excellence Scholarship	2019, 2020
UBC Faculty of Arts International Student Scholarship	2019, 2020, 2022