Updated March 2026

Muhua (黄牧华) Huang

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

I study how AI systems fundamentally differ from **human intelligence** and how they, as **social actors**, reshape **collaboration** and **collective decision-making**. My research combines **computational social science**, **psychometrics**, and **generative agent-based modelling (GABM)** to investigate AI's unique cognitive characteristics, uncover new insights about human behavior from AI representations, and understand how AI agents alter **interaction dynamics** within **groups and organizations**.

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

Education

Stanford Graduate School of Business 2025.9 – 2030.6
PhD in Organizational Behavior

The University of Chicago 2023.9 – 2025.6
MA in Computational Social Science; GPA: 3.9/4.0

The University of British Columbia (UBC) 2018.9 – 2023.5
BA in Computer Science & Psychology (Honours); GPA: 4.2/4.3

International AI Evaluation Programme Fellow 2026.1 – 2026.5

Summer Institute in Computational Social Science (SICSS) 2024.6 – 2024.7

Google Research Mentorship Program Scholar 2020.9 – 2021.1

Publications

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

[2] 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

[3] 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

[4] 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

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

[6] 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

[7] 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*. [Preprint]

Working Papers

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

[2] **Huang, M.**, & Evans, J. (In Press). Institutions as cached computation for resource-rational negotiation. *Behavioral and Brain Sciences*. DOI

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

[4] Kim, H., Yi, X., Yao, J., **Huang, M.**, Bak, J., Evans, J., & Xie, X. (2025). Research on Superalignment Should Advance Now with Parallel Optimization of Competence and Conformity. [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 Experience

Research Intern | Microsoft Research Asia 2024.7 – 2024.10
Advisor: Dr. Xiaoyuan Yi, Dr. Xing Xie

- Built a multi-agent simulation platform where hundreds of LLM agents interact, form social networks, and self-organize governance structures (e.g., constitutions) without external guidance.
- Grounded agent behavior in Schwartz's Theory of Basic Values and used Social Network Analysis to trace how value-diverse communities form, evolve, and stabilize over time.
- Discovered that moderate value diversity fosters creativity and stability in agent communities, while extreme heterogeneity induces instability — with implications for designing diverse human-AI teams.
- Investigated social emergence (unique language, cultural practices, institutions) in agent communities using GraphRAG and NLP.
- Contributed to interdisciplinary projects on AI values and alignment, including LLM value evaluation benchmarks (ACL 2025), psychometrically grounded trait elicitation, and a survey of superalignment.

Research Experience

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

AI as Alien Intelligence: Shared Vocabulary Masks Divergent Category Systems in Human and AI Collectives

- Engineered a large-scale multi-agent coordination experiment (∼375,000 AI trials with GPT-4o agents in network structures of N=2 to 50), replicating a human category-formation paradigm.
- Discovered "representational relativity": despite comparable behavioral performance, human and AI collectives resolve ambiguity in systematically different places ($r = -0.130$, $p < 0.001$).
- Showed AI labels are disproportionately visual-dominant and concrete, while human labels draw on richer multimodal experience — divergence driven by grounding modality, not training data.
- Found that AI agents strategically invent novel vocabulary to express distinctions human language cannot; forcing AI to use only human labels reduces performance by 12.7%.

Master's Thesis | University of Chicago 2023.9 – 2024.6
Advisor: Dr. James Evans

Designing LLM-Agents with Personalities: A Psychometric Approach

- Developed the first psychometrically validated framework for assigning Big Five personalities to LLM agents, with controllable, fine-grained design across multiple models (GPT-3.5, GPT-4, Claude 3).
- Validated that personality-assigned agents exhibit predictable behavioral differences in moral reasoning, risk-taking, and cooperation (personality explains 38–52% of variance in agent decisions).
- Collected data from 350 human participants for direct human–agent comparison, demonstrating potential for using personality-designed agents in behavioral research.
- Published in *Personality Science*; methodology enables designing AI agents for specific collaborative roles.

Research Assistant, Quantitative Psychology | University of British Columbia 2020.4 – 2023.8
Advisor: Dr. Victoria Savalei

Exploring the Factor Structure of the BFI-2 in Alternative Scale Formats

- Compared four personality scale formats using SEM; found that alternative formats (Expanded, Item-Specific) significantly improve reliability and model fit over Likert. Published in *Journal of Personality Assessment*.
- Honours thesis nominated for multiple Canadian national undergraduate research awards (Belkin, CPA).

SEM Fit Indices' Sensitivity to Omitted Crossloadings in CFA

- Built an R Shiny simulation app to examine fit index performance when CFA models are applied to data with EFA-like structure; contributed to manuscript published in *Psychological Methods*.

Research Assistant, Social Psychology | University of British Columbia 2020.8 – 2021.9
Advisor: Dr. Kristin Laurin

Socioeconomics and Imagined Future Study

- Applied ML (Random Forest, Neural Network) and NLP (Topic Modelling, LSTM, Word2Vec) to analyze 10,000+ youth essays, revealing how socioeconomic background shapes imagined futures. Published in *Journal of Social Issues*.
- Supervised 20 research assistants and built data processing protocols including a word distribution and performance tracking system.

Research Assistant, Human-Centered AI | University of British Columbia 2021.10 – 2022.9
Advisor: Dr. Christina Conati

Personalized Explainable AI for Intelligent Tutoring System Study

- Evaluated how personalizing AI explanations to individual cognitive profiles (perceptual speed, personality) affects learning outcomes, using eye-tracking and linear mixed models.
- Found that tailoring AI tutoring interfaces to users' cognitive abilities improves engagement and effectiveness — demonstrating the value of adaptive human-AI interaction design.

Professional Experience

Computational Social Science Workshop Coordinator | University of Chicago 2024.10 – 2025.6

- Organized the weekly MACSS Computational Social Science Workshop, featuring speakers on advanced computational methods and interdisciplinary research.

Deep Learning Textbook Editor | University of Chicago 2023.10 – 2024.10

- Proofread and edited "Thinking with Deep Learning" by Dr. James Evans; designed LaTeX templates for the textbook and for Journal of Social Computing.

Teaching Assistant | University of British Columbia 2021.01 – 2022.12

- TA for Human Computer Interaction Methods and Systematic Program Design; rated 4.7/5 by students and received three return offers.

Crisis Respondent | Kids Help Phone 2019.09 – 2020.08

- Provided over 230 hours of crisis counselling to 550+ individuals in severe mental distress.

Honors & Awards

Stanford Graduate School of Business Doctoral Fellowship	2025 – 2030
Stanford University EDGE Doctoral Fellowship ($12,800)	2025
UChicago Outstanding Thesis Award ($1,000)	2025
Microsoft Research Asia Outstanding Intern Award (Stars of Tomorrow Certificate)	2024
OpenAI Researcher Access Program ($3,500)	2024
SICSS-Beijing Merit Based Scholarship (¥3,600)	2024
UChicago Computational Social Science Research Poster Competition (Honorable Mention, 2nd Place)	2024
UChicago Quadrangle Research Scholarship ($80,000)	2023, 2024
UChicago Social Sciences Promise & Merit Scholarship ($10,000)	2023, 2024
Quinn Research Assistantship Award ($8,300)	2021
UBC International Community Achievement Award ($5,000)	2021
UBC Trek Excellence Scholarship ($2,000)	2019, 2020
UBC Faculty of Arts International Student Scholarship ($8,800)	2019, 2020, 2022