Xiao Ma

Alignment Researcher | Meta Superintelligence Labs

Palo Alto, CA | maxiao.info | Google Scholar | alphaXiv

Alignment researcher working across post-training, reinforcement learning, and scalable safety evaluations for frontier models and agents. Experience spans method invention, empirical validation, and deployment at Meta and Google DeepMind, grounded in research on safety alignment and human trust.

Research interests: Multi-agent alignment and emerging hybrid human-agent network dynamics, particularly coordination and social engineering risks.

Professional Experience

Apr 2026-present **Meta Superintelligence Labs**, Alignment Researcher

- Recently led alignment of the image and video planner agent, incorporating training data and reward improvements and running alignment training jobs.
- Formalized safety evaluations and integrated the full safety evaluation suite into scalable infrastructure.
- Critical contributor to the July 2026 launch of Muse Image.

2024-Apr 2026 **Google DeepMind**, Staff Software Engineer, *GenAI*

- Co-invented and empirically validated **RLGF (grading feedback)**, an RL method using reasoning and self-improvement to improve model safety without sacrificing quality. Brought it into production in late 2024 to unblock Gemini 2.5 Pro and beyond.
- Scaled RLGF across Gemini through technical documentation, a playbook, and a summit, supporting safety, mathematical reasoning and proof generation (IMO gold-medal model), system instruction following, and agentic capabilities.

2019-2024 **Google Research**, Senior Software Engineer

- Designed launch-critical Bard/Gemini safety evaluations for 3+ launches and contributed synthetic-data mitigations across 5+ launches.
- Built and launched an AI-assisted video review system for YouTube moderators across 3+ launches, significantly improving reviewer productivity.

2018, 2017 **Facebook**, PhD Research Intern, *Core Data Science*
Graph-based research on social trust and group engagement.

2016 **Airbnb**, PhD Research Intern
Large-scale analysis of user and host trust.

Education

2014-2019 **Cornell Tech, Cornell University**
Ph.D., Information Science, Minor in Computer Science
Thesis: *Networked Trust: Computational Understanding of Interpersonal Trust Online*
Committee: Mor Naaman (advisor), Karen Levy, Serge Belongie, Jeff Hancock

2010-2014 **Peking University**
Bachelor of Science, Microelectronics
Focused on chip design and manufacturing.

Selected Publications

2025 **Safety Alignment Should Be Made More Than Just a Few Tokens Deep.**
Xiangyu Qi, Ashwinee Panda, Kaifeng Lyu, **Xiao Ma**, Subhrajit Roy, Ahmad Beirami, Prateek Mittal, Peter Henderson.
ICLR 2025
Oral Presentation, Outstanding Paper Award

Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities.
Gemini Team.
Technical report, 2025

2023 **Gemini: A Family of Highly Capable Multimodal Models.**
Gemini Team.
Technical report, 2023

Let's Do a Thought Experiment: Using Counterfactuals to Improve Moral Reasoning.
Xiao Ma, Swaroop Mishra, Ahmad Beirami, Alex Beutel, Jilin Chen.
ICML Neural Conversational AI Workshop, 2023

A Mixed-Methods Approach to Understanding User Trust after Voice Assistant Failures.
Amanda Baughan, Xuezhi Wang, Ariel Liu, Allison Mercurio, Jilin Chen, **Xiao Ma**.
CHI 2023
Honorable Mention for Best Paper

2022 **A Human-ML Collaboration Framework for Improving Video Content Reviews.**
Meghana Deodhar, **Xiao Ma**, Yixin Cai, Alex Koes, Alex Beutel, Jilin Chen.
CIKM Human-in-the-Loop Data Curation Workshop, 2022
Best Paper

2019 **AI-Mediated Communication: How the Perception that Profile Text was Written by AI Affects Trustworthiness.**
Maurice Jakesch, Megan French, **Xiao Ma**, Jeffrey T Hancock, Mor Naaman.
CHI 2019
Full publication list: alphaXiv.

Mentoring & Service

Mentored Xiangyu Qi (Princeton, 2024) and Amanda Baughan (University of Washington, 2021).

Reviewer for COLM, ICLR, EMNLP, ICML, NeurIPS, FAccT, CSCW, CHI, CVPR, and ICWSM.