

Ram Potham

ram.potham@gmail.com • github.com/rapturt9 • linkedin.com/in/rampotham

Publications

MAEBE: Multi-Agent Emergent Behavior Framework

Developed framework for analyzing emergent behaviors in multi-agent systems, focusing on safety and alignment in complex AI environments

Accepted: HICSS Trustworthy AI Track, ICML Multi-agent Systems Workshop

Evaluating LLM Agent Adherence to Hierarchical Safety Principles

Lightweight benchmark using gridworlds for evaluating LLM agent ability to uphold high-level safety principles when faced with conflicting lower-level instructions

Accepted: ICML Technical AI Governance Workshop (**Oral Presentation**)

Research Experience

Redwood Research

Astra Fellow

January 2026 - Current

- High-stakes control projects

MIT Algorithmic Alignment Lab

AI Safety Researcher

June 2025 - October 2025

- Collaborated with **Dylan Hadfield-Menell** and **Stewart Slocum (Anthropic Fellow)** on character science.
- Developed open source **character training** pipeline increasing consistency by 15% over prompted baseline

Alignment Research Fellowship (AI Safety Global Society)

Mentor

April 2025 - July 2025

- Mentored cohort of researchers on empirical alignment techniques and safety evaluations with **ARENA**

Carnegie Mellon - Chimps Lab

AI/HCI Researcher

May 2022 - April 2023

- Developed crowd-auditing framework (**WeAudit**) to identify robustness failures in AI models, focusing on bias
-

Industry Experience

Watertight AI

Technical Staff

November 2025 - December 2025

- Building AI safety tech for major labs

Sitewiz (exited to GAIN)

Founder / CEO

November 2023 - May 2025

- Built autonomous AI agents for web development, analytics, and UI/UX automation for leading CRO agencies
-

Technical Skills

- **Alignment Research:** Empirical alignment, alignment stress-testing, safety evaluations, multi-agent systems
 - **ML Engineering:** PyTorch, TensorFlow, LLM fine-tuning, reinforcement learning, transformer architectures
 - **Production Systems:** Python, AWS, distributed systems, MLOps, autonomous agent deployment
-

Education

CARNEGIE MELLON UNIVERSITY, School of Computer Science

December 2024

Bachelor of Science in Artificial Intelligence | QPA 3.69

Recognition

- **Winner, AI Safety Action Competition** | *European Network for AI Safety*, 2025
- **USAMO qualifier** | *The United States of America Mathematical Olympiad*