Ram Potham

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

Research Experience

CBAI Fellow (MIT / Harvard)

AI Safety Fellow — June 2025 - Current

- Working with Dylan Hadfield-Menell (MIT Assistant Professor) and Stewart Slocum (MIT CSAIL PhD)
- Building open-source character training pipeline

Alignment Research Fellowship (AI Safety Global Society)

Mentor — April 2025 - Current

- Teaching ARENA curriculum and guiding discussions on AI Safety fundamentals
- Mentoring researchers on alignment techniques and safety evaluation methods

Carnegie Mellon - Chimps Lab

AI/HCI Researcher — May 2022 - April 2023

- Developed a crowd-auditing framework (WeAudit) to identify robustness failures in AI models, focusing on bias
- Contributed to human-in-the-loop AI evaluation, researching how AI misclassifications impact decision-making

Publications & Preprints

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 to Technical AI Governance Workshop at ICML 2025 (Oral Presentation)

Corrigibility as a Singular Target: A Vision for Inherently Reliable Foundation Models

Vision paper addressing corrigibility enhancement in AI systems to reduce loss of control scenarios, outlining empirical approaches for training corrigible foundation models

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 to Multi-Agent Systems Workshop at ICML 2025 (Poster)

Industry Experience

Sitewiz

Founder / CEO / CTO — November 2023 - May 2025

- Developed autonomous AI agents for, analytics and UI/UX automation, ensuring alignment with business KPIs
- Reduced reliance on human oversight by 90%, demonstrating safe deployment of automated insights

Harness

Machine Learning Intern — May 2022 - August 2022

- Developed time-series forecasting models for cloud cost projections
- Implemented anomaly detection using adaptive error thresholds to flag billing anomalies

Education

CARNEGIE MELLON UNIVERSITY, School of Computer Science — December 2024

Bachelor of Science in Artificial Intelligence | University Honors

Selected Coursework: Autonomous Agents, Natural Language Processing, Visual Language and Recognition

Technical Skills

- **AI Safety:** Corrigibility, Safety Case, Robustness Testing, Evaluations, Guardrails
- **ML & Cloud:** Python (TensorFlow, PyTorch), AWS (Serverless, Lambda, S3), Multi-agent Systems, Fine-Tuning
- **Research & Strategy:** Empirical Alignment Research, Systems Mindset, First Principles Mindset, Scrum