Akash Kundu

Kolkata, West Bengal, India
akashkundu2xx4@gmail.com — LinkedIn — Google Scholar — Website

EDUCATION

Heritage Institute of Technology, Kolkata, India — Oct 2022 – Jun 2026
Bachelor of Technology in Computer Science and Engineering (GPA: 8.75/10) — Final Year GPA: 9.33/10

PUBLICATIONS

- Esben Kran*, Hieu Minh Nguyen*, **Akash Kundu***, Sami Jawhar*, Jinsuk Park*, and Mateusz Maria Jurewicz (2025). "DarkBench: Benchmarking Dark Patterns in Large Language Models". In: ***The Thirteenth International Conference on Learning Representations***. URL: https://openreview.net/forum?id=odjMSBSWRt
- **Akash Kundu**, Emanuel Tewolde, Ratip Emin Berker, Samuel F. Brown, and Vincent Conitzer (2026). "Do LLMs Take Care of Their Own? Similarity Signals Can Induce Cooperation". In: Accepted at **AI4GOOD Workshop @ ICML 2026**; under review at **NeurIPS 2026**
- K. Enevoldsen, I. Chung, I. Kerboua, et al. (incl. **A. Kundu**). "MMTEB: Massive Multilingual Text Embedding Benchmark." *ICLR 2025.* https://openreview.net/forum?id=zl3pfz4VCV
- C. Smith, M. Abdulhai, M. Diaz, et al. (incl. **A. Kundu**). "Evaluating Generalization Capabilities of LLM-Based Agents in Mixed-Motive Scenarios Using Concordia." *NeurIPS 2025 Datasets & Benchmarks Track.* https://openreview.net/forum?id=yG4Fj0voJZ
- **Akash Kundu**, Adrianna Tan, Theodora Skeadas, Rumman Chowdhury, and Sarah Amos (2025). "Red Teaming for Trust: Evaluating Multicultural and Multilingual AI Systems in Asia-Pacific". In: ***ICLR 2025 Workshop on Building Trust in Language Models and Applications***. URL: https://openreview.net/forum?id=SPlhZYuH9e
- Reva Schwartz, Rumman Chowdhury, **Akash Kundu**, Heather Frase, Marzieh Fadaee, Tom David, Gabriella Waters, Afaf Taik, Morgan Briggs, Patrick Hall, Shomik Jain, Kyra Yee, Spencer Thomas, Sundeep Bhandari, Paul Duncan, Andrew Thompson, Maya Carlyle, Qinghua Lu, Matthew Holmes, and Theodora Skeadas (2025). *Reality Check: A New Evaluation Ecosystem Is Necessary to Understand AI's Real World Effects.* arXiv: 2505.18893 [cs.CY]. URL: https://arxiv.org/abs/2505.18893
- **Akash Kundu** and Rishika Goswami (2025). "AI Through the Human Lens: Investigating Cognitive Theories in Machine Psychology". In: ***IJCNLP-AACL 2025 Student Research Workshop*** *(Direct Submission)*
- Arnab Sengupta, **Kundu, Akash**, and Aishi Mukhopadhyay (2024). "An Approach to Detect and Classify Potentially Suspicious Activity from Real-Time Log Data using Anomaly Detection Methods". In: ***2024 3rd International Conference for Innovation in Technology (INOCON)***, pp. 1–9. DOI: 10.1109/INOCON60754.2024.10511679
- Wm. Matthew Kennedy, Rumman Chowdhury, Reva Schwartz, Cigdem Patlak, Jack Hagen, Blake Chambers, **Kundu, Akash**, Liam Baisley, Aauysh Dhanotiya, Jayraj Dave, Sukanya Moorthy, and Theodora Skeadas (2025). "Ask What Your Country Can Do For You: Towards a Public Red Teaming Model". In: ***CAMLIS Red Workshop***

SELECTED RESEARCH

Do LLMs Take Care of Their Own? Similarity Signals Can Induce Cooperation ICML AI4GOOD Workshop 2026

- First-authored (with **V. Conitzer**) a study showing that **continuous similarity signals can induce cooperation** in LLM agents across social dilemmas, even when defection is the dominant strategy; under review at **NeurIPS 2026**.
- Found that cooperation tracks **perceived** similarity over measured similarity: models only ˜46% alike by external benchmarks judge themselves 92–94% similar when reasoning about an opponent directly, and cooperate where the measured score predicts defection.
- Built the evaluation as an **Inspect-compatible benchmark suite** spanning six social-dilemma games for community reuse.

DarkBench: Benchmarking Dark Patterns in Large Language Models ICLR 2025 (Oral)

- Developed DarkBench, a novel benchmark to evaluate dark patterns in LLMs, with 660 adversarial prompts across six categories.
- Evaluated 14 open-source and proprietary models, uncovering widespread dark patterns and ethical gaps.
- Accepted as an oral (top 1.8%) at **ICLR 2025**; preliminary version at **AAAI 2025 DATASAFE workshop**.

Red Teaming for Trust: Evaluating Multicultural and Multilingual AI Systems in Asia-Pacific ICLR Building Trust Workshop 2025

- Authored comprehensive research on AI safety, introducing the first multicultural and multilingual AI Safety Red Teaming Challenge in the Asia-Pacific region.
- Designed and conducted a large-scale study involving 54 participants from 9 countries, assessing Large Language Models (LLMs) across diverse cultural and linguistic contexts.
- Authored the full research paper, including methodology, findings, and recommendations.

Massive Multilingual Text Embedding Benchmark (MMTEB) ICLR 2025

- Co-authored MMTEB, expanding MTEB to 500+ evaluation tasks across 1,000+ languages.
- Earned co-authorship through open-source contributions, enabling methodological innovations that reduced computational demands and improved benchmarking efficiency.
- Paper accepted at **ICLR 2025**, with unanimous reviewer appreciation for the benchmark's depth, innovation, and potential impact on the field.

Evaluating Generalization of LLM Agents in Mixed-Motive Scenarios using Concordia NeurIPS 2025 Datasets and Benchmarks

- Part of the winning team in the hackathon that preceded the NeurIPS Concordia Contest, 2024 organized by **Google DeepMind** and the **Cooperative AI Foundation**.
- Our agent consistently ranked among the top performers on the leaderboard, leading to an invitation as co-authors on the resulting paper.
- The paper presents a benchmark and evaluation framework for multi-agent alignment under conflicting objectives, offering insights into robustness and coordination in complex environments.

INVITED TALKS

IIT Delhi · Secure AI Futures Lab Jun. 20, 2026
Getting Into AI Safety Research: Opportunities and Pathways

Global South AI Safety Hackathon · Apart Research Jun. 18, 2026
Dark Patterns in Large Language Models [video]

Cooperative AI Research Fellowship Showcase · Cape Town · selected as 1 of 3 fellows (of 10) Apr. 21, 2026
Similarity as a Signal: Do AI Agents Cooperate More When They Know They're Alike? [video]

University of Cape Town · ShockLab Seminar Series Apr. 01, 2026
Similarity as a Signal: Do AI Agents Cooperate More When They Know They're Alike?

Stellenbosch University · Policy Innovation Lab · AI Safety Research Workshop Mar. 27, 2026
Similarity as a Signal: Do AI Agents Cooperate More When They Know They're Alike?

PROFESSIONAL SERVICE

Organizing Committee & Reviewer, AI4GOOD Workshop @ ICML 2026 2026

- Supported workshop operations, including reviewer recruitment and desk-rejection screening, and reviewed submissions on alignment faking and multi-agent cooperation.

Judge, AI Manipulation Hackathon, Apart Research 2026
Reviewer, ICLR 2025 BuildingTrust Workshop 2025

EXPERIENCE

Research Fellow, Cooperative AI Research Fellowship Feb 2026 – Present *Cape Town, South Africa*

- Selected as one of 10 fellows from a global pool of over 1,000 applicants ($< 1\%$ acceptance rate) to participate in a selective residency program focused on AI safety.
- Conducting empirical research on the foundations of multi-agent risks and cooperative game theory under the mentorship of Prof. Vincent Conitzer.
- Have given in-person talks about my research at Stellenbosch University and the University of Cape Town during my fellowship.
- Awarded a funded extension to continue research through August 2026, one of two fellows selected for the extension.

Research Collaborator, FAR.AI Sep 2025 – Dec 2025, Berkeley, CA

- Collaborated with FAR.AI to develop a toolkit that empowers expert frontier model red-teaming.

Red Teaming Data Scientist, Humane Intelligence Jan 2025 – Jul 2025, Remote

- Contributed to the Red Teaming Evaluations team by analyzing post-event red-teaming data and preparing comprehensive reports.
- Developed an Auto-Red Teaming LLM that autonomously generated exploit-style prompts to test other LLMs for multilingual and sociocultural safety vulnerabilities using LoRA fine-tuning and causal language modeling.
- One of the core contributors to the Singapore AI Safety Red Teaming Challenge in collaboration with the Singapore Government (IMDA).
- Co-authored multiple papers on Red Teaming, Science of Evals, and LLM biases.

Research Fellow, Apart Research Jan 2024 – Mar 2025, Remote

- Co-authored a paper researching cross-lingual capabilities of Safety Evaluations Benchmark.
- Co-authored a paper evaluating Dark Patterns in State-of-the-Art LLMs, including the development of a 600+ dataset to benchmark dark patterns across 6 distinct categories.
- One paper accepted at **ICLR 2025 (Oral)** and **AAAI 2025**.

Research Intern, Lionheart Ventures Jun 2024 – Sep 2024, San Francisco, CA

- Developed a comprehensive framework to evaluate AI-induced systemic risks across 10+ portfolio companies.
- Performed cluster analysis to categorize and identify patterns among 600+ distinct AI-related threats.
- Conducted Monte Carlo simulations to assign risk weights to identified threats, using methodologies similar to MIT's AI Risk Initiative.
- Integrated framework into internal due diligence processes for funding evaluations.

VOLUNTEERING

AI/ML Lead / Tech Lead, Google Developers Student Club Jul 2023 – Jul 2025

- Led the club's technical domain, managing a team of 14 domain mentors across web, Android, and ML to support a community of over *1800* active members.
- Designed and led hands-on ML coding sessions, engaging over *150* active participants and facilitating technical skill development.
- Launched the institute's inaugural ML Hackathon, successfully hosting over *300* participants.

Volunteer ML Engineer, Omdena Feb 2023 – Oct 2023, Remote

- Developed a deepfake detection system for women safety in Germany.
- Estimated out-of-pocket lung cancer costs for patients in the US with a confidence interval of 10%.
- Created a matching algorithm for hostel roommates based on personality traits in Egypt.
- Fine-tuned Vicuna 13B for a mental health chatbot supporting English and Swahili, aimed at assisting people in Tanzania.

CERTIFICATION

Introduction to Cooperative AI, Cooperative AI Foundation 2025

AI Safety Fundamentals — Alignment Course, BlueDot Impact 2024

ACHIEVEMENTS

- Recipient of a $1,000 Travel Grant for the AI4GOOD Workshop @ ICML 2026.
- Accepted to the Cooperative AI Summer School 2026, Toronto.
- Past Fellow at Supervised Program for Alignment Research, collaborating with FAR.AI.
- Past Fellow at Whitebox Research studying mechanistic interpretability.
- Presented solution at the Concordia Workshop of NeurIPS 2024 upon invitation.
- Received $1,100 from Google DeepMind as Researcher credits for the NeurIPS Concordia Competition, 2024.
- Secured 1st position in the Concordia Hackathon hosted by Apart Research and Cooperative AI Foundation in Sep 2024.
- Won the Spectral Syntax Bounty worth $2,000 in the Web3 x AI Hackathon by Encode Club in July 2024.
- Accepted in CaMLAB V4, an ARENA equivalent ML BootCamp organized by Cambridge AI Safety Hub.
- Winner of LLM Evaluations Hackathon (Nov '23) and AI Security Evaluations Hackathon (May '24) organized by Apart Research.