Akash Agrawal

Email Website LinkedIn Google Scholar

Education

University of Oxford *Sept 2023 – Sept 2024*
Master of Science, Advanced Computer Science

Indian Institute of Technology, Delhi *July 2019 – May 2023*
Bachelor of Technology, Electrical Engineering

Publications

The Multi-Agent Off-Switch Game [link] 2025
Akash Agrawal*, Soroush Ebadian*, Lewis Hammond.
AAMAS '26 (**Oral**). Also in AAAI '26 Workshops on FAST (**Oral**) TrustAgent.

Robust Policy Design in Agent-Based Simulators using Adversarial RL [link] 2025
Akash Agrawal, Joel Dyer, Aldo Glielmo, Michael Wooldridge.
AAAI '25 Workshop on Multi-Agent AI in the Real World (**Best Paper Award** and **Oral**).

Manuscripts Under Review

ϵ−Nash Equilibria for Observation Space Attack Games 2026
Dominik B. Zurcher, Brandon Kaplowitz, **Akash Agrawal**, Tala Jafari, Christian Schroeder de Witt, Paul Goldberg.
Under Review at ICML 2026.

Achievements and Awards

- **Cooperative AI PhD Fellowship**: Awarded fellowship for doctoral research (**∼5% acceptance rate**).
- **Best Paper Award**: Awarded for work on robust RL in agent-based models at **AAAI MARW 2025**.
- **Informatics Olympiad 2018**: Qualified for the training camp for IOI, ranking **10th in India**.
- **MATS Scholar**: Selected for prestigious AI Safety Research Fellowship (**∼2% acceptance rate**) for summer 2025.
- **Funding**: Awarded **∼$93,000 in funding** to work on Cooperative AI through AI Safety Support.
- **Cooperative AI Summer School 2025**: Selective summer school (**∼13% acceptance**) for researchers (funded).
- **Philosophy Olympiad**: Ranked **4th nationwide** in the first round of the International Philosophy Olympiad 2018.
- **Department Change**: Granted department change at IIT Delhi for being in **top 5% students** in first year.
- **JEE Mains**: Ranked in the **99.8th percentile** (out of 1.3 million) in JEE Mains 2019 all over India.
- **Winner, ICPC for Schools**: Won national-level programming contest across all schools in India (2018).
- **APIO 2018**: Participated in the Asia Pacific Informatics Olympiad as member of Indian cohort.

Research Projects

Foundations of Multi-Agent Shutdownability *June 2025 - Present*
ML Alignment & Theory Scholars Lewis Hammond

- Developed a **game-theoretic framework** for shutdown-preservation incentives in multi-agent AI systems.
- Proved **conditions where shutdownability composes** and when **collective shutdown-resistance emerges**.
- Identified **belief divergence** and **joint-action utilities** as key structural drivers of shutdown-resistance.
- Wrote co-first-authored paper *The Multi-Agent Off-Switch Game*, which was **accepted to AAMAS 2026 for oral presentation** and to **two AAAI 2026 workshops: FAST (Oral) and TrustAgent**.

Robust Policy Design in Agent-Based Models *June 2024 – Present*
University of Oxford Prof. Michael Wooldridge, Dr. Joel Dyer, Dr. Aldo Glielmo

- Showed **RL-based policy design in ABMs** is **sensitive to misspecification** of environment parameters.
- Identified **two adversarial-training schemes** to improve **sim-to-real robustness** of ABM-derived policies.
- Empirically showed adversarially trained **policies improve** performance under **both seen and unseen shifts**.
- Wrote first-authored paper *Robust Policy Design in Agent-Based Simulators using Adversarial Reinforcement Learning* which received the **Best Paper Award at AAAI 2025 MARW Workshop**.

Computing Equilibria in Illusory Attack Games *May 2025 – Present*
University of Oxford Dr. Christian Schroeder de Witt

- Formulated illusory **observation-space attacks** as a two-player zero-sum security game with partial observability.
- Proved robust defence can **require history-dependent policies**, even when underlying environment is Markov.

- Adapted equilibrium algorithm to compute **approximate robust strategies** and are implementing experiments.

Hardness of Socially Manipulating Elections *May 2022 – August 2022*
Washington University in St. Louis Prof. Yevgeniy Vorobeychik

- Formalized **manipulation** of **issue-based elections** on networks with **independent cascade of misinformation**.
- Reduced the social election manipulation problem to the positive–negative influence-maximization problem.
- Proved that **approximating** the optimal attack is **intractable** by reducing from **+/- partial set cover**.

Diversity in Rank Aggregation *Dec 2021 – May 2023*
Indian Institute of Technology, Delhi Prof. Abhijnan Chakraborty

- Formulated the problem of selecting diverse set of rankings from a voting profile to reduce coordinated influence.
- Designed a fixed-parameter tractable algorithm to combine multiple voting rules to generate diverse ranking sets.

Fault-Tolerant Data Structures for Min-Cuts *Apr 2021 - Aug 2021*
Indian Institute of Technology, Delhi Prof. Keerti Choudhary

- Studied data structures for s–t min-cut queries under deletion of up to r edges in a dynamic graph.
- Developed an $O(mn)$-space structure to answer all $r = 2$ failures scenario in $O(1)$ query time.
- Analysed candidate approaches for $r = 3$ and proved why analogous efficient structures cannot be obtained.

Course Projects

Evaluating Variational Continual Learning with a Laplace Prior *Mar 2024 – Apr 2024*

- Implemented VCL with a Laplacian prior over network weights and empirically improved performance over the Gaussian-prior baseline on Permuted and Split MNIST (in Uncertainty in Deep Learning with Prof Yarin Gal).

Learnable Message Aggregation for Link Prediction in Temporal Graphs *Mar 2024 – Apr 2024*

- Implemented TGNs with attention- and RNN-based message aggregation and improved link-prediction accuracy on the Wikipedia dataset using attention (in Geometric Deep Learning with Prof. Michael Bronstein).

Normative Disagreement as a Challenge for Cooperative AI *Apr 2023 – May 2023*

- Surveyed work on normative disagreement in multi-agent settings and presented a synthesis of key ideas and open problems in a course talk (in Cooperative AI and RL Independent Study course with Prof. Rohit Vaish).

Fairness in the Multi-Agent Multi-Armed Bandit Problem *Mar 2022 – Apr 2022*

- Reviewed regret and fairness notions in multi-agent bandits and analysed trade-offs, presenting conclusions and open directions in a report and talk (in Computational Social Choice with Prof. Rohit Vaish).

Selected Coursework

Math/CS: Cooperative Game Theory, Computational Game Theory, Computational Social Choice, Cooperative AI and Reinforcement Learning, Computational Learning Theory, Randomized Algorithms, Probabilistic Deep Learning, Geometric Deep Learning, Optimization for Machine Learning, Advanced Graph Data Structures.

Policy/Philosophy: AI Ethics and Safety, Law and Computer Science, Science Technology and Human Development, Social Science Approaches to Development, Inclusive Innovation, Critical Philosophy of Race, Introduction to Ethics

Leadership Experience

Head, Student Curriculum Review Committee, IIT Delhi *Jun '22 - May '23*

- Led a 40+ student committee to design and run IIT Delhi's curriculum-feedback process, synthesising student input into proposals for the institute-wide curriculum review.

Founder and President, Tech Club, DPS Ruby Park *Sep '17 - Mar '19*

- Founded the led the school's Tech Club and tech fest (biggest in east India), leading a tiered student team to run popular workshops and competitions in programming, development, and robotics.

Founder, Effective Altruism Society, IIT Delhi *Aug '22 - May '23*

- Founded one of India's first Effective Altruism societies; held talks, reading groups, and workshops on doing impact.

Extra Curricular

Music: Played guitar in various events and concerts. Completed Grade 4 certification from University of West London.
Debating: Participated in national and international debating competitions. Won prizes for adjudication.

Volunteering

Girls Who Code: Taught the basics of computer science to underprivileged 6th-grade students.
Pratham Education: Created educational content to improve reading abilities of underprivileged children.