# Akbir Khan

akbir.dev

## Education

2021-2024 Ph.D. in Computer Science, University College London<br>
Advised by Ed Grefenstette & Tim Rocktäschel

2017-2018 MPhil in Advanced Computer Science, *with distinction*, University of Cambridge

2013-2017 MSci in Mathematics with Physics, *with 1st class honours*, University College London

2015-2016 Exchange student, *as Mathematics Specialist*, University of Toronto

## Experience

2023- Research Analyst at Cooperative AI Foundation, grant-making and field-building to mitigate risks posed by multi-polar AI outcomes

2021-2023 Senior Applied Researcher at Tractable AI, built OCR pipeline which generates £8M ARR

2017-2020 CRO at Spherical Defence, built Seq2seq models for anomalous web traffic detection

2016 Four month internship at Deutsche Bank as a Software Engineer

2015 Research Intern at the Quantum Optics and Laser Group, Imperial College London

## Selected Publications

2024 Debating with More Persuasive LLMs Leads to More Truthful Answers — **A Khan**, J Hughes, D Valentine, L Ruis, K Sachan, A Radhakrishnan, E Grefenstette, S Bowman, T Rocktäschel & E Perez. **Oral [top 1.5%]** at *International Conference on Machine Learning (ICML)*

2024 Scaling Opponent Shaping to High Dimensional Games — **A Khan**, T Willi, N Kwan, A Tachetti, C Lu, T Rocktäschel, E Grefenstette & J Foerstor. Oral at *The Autonomous Agents and Multi-Agent Systems (AAMAS)*

2023 The Goldilocks of Pragmatic Understanding: Fine-Tuning Strategy Matters for Implicature Resolution by LLMs — L Ruis, **A Khan**, S Biderman, S Hooker, T Rocktäschel, & E Grefenstette. **Spotlight [top 3%]** at *The Neural Information Processing Systems (NeurIPS)*

2023 MAESTRO: Open-Ended Environment Design for Multi-Agent Reinforcement Learning — M Samvelyan, **A Khan**, M Dennis, M Jiang, J Parker-Holder, JN Foerster, R Raileanu, T Rocktäschel. Accepted at *International Conference on Learning Representations (ICLR)*

## Awards

2024 SuperAlignment Fellowship, OpenAI

2023 Astra Fellowship, Redwood Research

2020 Foundational Artificial Intelligence Scholarship, ESPRC

## Technical Projects

Deep Equilibrium Models, a Haiku implementation of the NeurIPS 2019 paper, an implicit-depth differentiable architecture that simulates an infinitely deep network

Bad Flamingo, gamified collection of adversarial training examples; awarded 1st place at Hack Cambridge Ternary

Skills: Python [PyTorch, JAX (contributor), Pandas, Scikit-learn], Docker, GoLang