

Akbir Khan

[akbir.dev](#)

Education

2021-2024	Ph.D. in Foundational Artificial Intelligence, University College London Advised by Ed Grefenstette & Tim Rocktäschel
2017-2018	MPhil in Advanced Computer Science, <i>with distinction</i> , University of Cambridge
2013-2017	MSci in Mathematics with Physics, <i>with 1st class honours</i> , University College London
2015-2016	Exchange student, <i>as Mathematics Specialist</i> , University of Toronto

Experience

2023-	Research Analyst at Cooperative AI Foundation , grant-making and encouraging research to mitigate risks posed by multi-polar AI outcomes
2021-2023	Senior Applied Researcher at Tractable AI , built OCR service which now generates £8 million in annual revenue
2017-2020	Chief Research Officer at Spherical Defence , developed Seq2seq models for web application firewalls; raised a \$2 million seed round
2016	Software Engineer Internship at Deutsche Bank
2015	Research Intern at the Quantum Optics and Laser Group , Imperial College London

Selected Publications

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. Accepted at *The Forty-first International Conference on Machine Learning*

Scaling Opponent Shaping to High Dimensional Games - **A Khan**, T Willi, N Kwan, A Tachetti, C Lu, T Rocktäschel, E Grefenstette & J Foerster. Oral at *The 23rd International Conference on Autonomous Agents and Multi-Agent Systems*

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 at *Thirty-seventh Conference on Neural Information Processing Systems*

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 *The 10th International Conference on Learning Representations*

Technical Projects & Skills

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

[Bad Flamingo](#), a gamified data collection of sketches for adversarial machine learning. Awarded 1st Prize at the University of Cambridge Ternary Hackathon

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