Joe Benton

+44 (0) 7964 070 158 | joe.benton127@gmail.com | St Peter's College, New Inn Hall Street, Oxford, OX1 2DL

Education

D.Phil. in Statistics 2021 – 2023 (Expected)
Department of Statistics, University of Oxford
Thesis title: *Generative Modelling: Theory and Applications*, supervised by Arnaud Doucet and George Deligiannidis

B.A. with M.Math. in Mathematics 2017 – 2021
Trinity College, University of Cambridge
Graduated with a Distinction (6th out of c. 250 in year)
Thesis title: *Activated Random Walks*, supervised by Perla Sousi

Work Experience

Machine Learning Researcher | *Redwood Research* Winter 2022 – 2023
Developed and studied causality and fine-tuning based interpretability methods, with applications to mechanistic anomaly detection. Supervised a month-long intern project aiming to automate interpretability techniques.

Research Intern | *Alignment Research Center* Spring 2023
Worked on formalizing heuristic arguments, finding efficient heurstic estimators for sparse covariance propagation.

Research Assistant | *Center for Human-Compatible Artificial Intelligence, UC Berkeley* Summer 2021
Built on the PAIRED algorithm for unsupervised environment design to incorporate human feedback with the aim of speeding up and simplifying the training process. Supervised by Michael Dennis.

Research Assistant | *DPMMS, University of Cambridge* Summer 2020
Adapted methods from algebraic geometry for the calculation of characteristic numbers to produce more direct computations of intersection numbers on the projective plane and torus. Supervised by Dhruv Ranganathan.

Volunteering

Research Mentor | *Supervised Program for Alignment Research* February 2022 – Present
Mentored a student research project on decoding sparse feature representations for neural network interpretability.

Strategic Advisor | *AI Safety Hub* March 2022 – Present
Advised an AI safety outreach and mentoring organization, and manged their AI Safety Fundamentals program.

Trustee, Cofounder | *Raise: A Celebration of Giving* February 2018 – Present
Trustee and co-founder of Raise, a student charity initiative raising over £460,000 for the Against Malaria Foundation.

Awards

International Mathematics Olympiad (1 Gold – 7th out of 615, 3 Silver) 2014 – 2017
International Olympiad in Informatics (1 Gold – 6th out of 304, 1 Silver, 1 Bronze) 2015 – 2017
Romanian Masters in Mathematics (3 Gold – Best record of any competitor) 2015 – 2017

Publications

Error Bounds for Flow Matching Methods. **Joe Benton**, George Deligiannidis, Arnaud Doucet. arXiv preprint, *arXiv:2305.16860*

From Denoising Diffusions to Denoising Markov Models. **Joe Benton**, Yuyang Shi, Valentin De Bortoli, George Deligiannidis, Arnaud Doucet. arXiv preprint, *arXiv:2211.03595*

Alpha-divergence Variational Inference Meets Importance Weighted Auto-Encoders: Methodology and Asymptotics. Kamélia Daudel, **Joe Benton***, Yuyang Shi*, Arnaud Doucet. *arXiv preprint, arXiv:2210.06226*

Polysemanticity and Capacity in Neural Networks. Adam Scherlis, Kshitij Sachan, Adam S. Jermyn, **Joe Benton**, Buck Shlegeris. *arXiv preprint, arXiv:2210.01892*

A Continuous Time Framework for Discrete Denoising Models. Andrew Campbell, **Joe Benton**, Valentin De Bortoli, Tom Rainforth, George Deligiannidis, *Arnaud Doucet. Advances in Neural Information Processing Systems, 2022*

Reviewing

TMLR, ICML 2023 Workshop Frontiers4LCD, Cooperative AI Foundation