Benjamin Berczi

Machine Learning Scientist

benji.berczi@gmail.com
benjibrcz
benjamin-berczi-b7bb0a133
0000-0001-9390-6124

Experience

Research Scholar | ML Alignment & Theory Scholars

January 2026 – Berkeley, USA

- AI safety research fellowship working with Cozmin Ududec of UK AISI
- Project working on elicitation and evaluations of LLMs

Research Scientist | Independent

June 2025 – London, UK

- Independent research project in AI safety in collaboration with Google DeepMind researcher Jonathan Richens
- Interpreting and steering the model internals of a goal-conditioned reinforcement learning agent to study how the agent represents its world model and goals, drawing from causal inference theory

Machine Learning Scientist | Depop

January 2024 – January 2026 London, UK

- Developed ML powered recommender system with PyTorch relying on a large scale ETL pipeline
- Gained experience using tools like Spark, Databricks, Airflow and AWS services

Research Fellow | AI Safety Camp

January 2024 – June 2024 London, UK

- Research in AI safety working on the interpretability of maze solving transformers
- Successfully implemented activation steering to control transformers and understand their representations

Data Scientist | Faculty AI

June 2023 – Aug 2023 London, UK

- Intensive data science course and internship, studied topics including: supervised learning, unsupervised learning, neural networks, decision trees, reinforcement learning, AI safety
- Developed a recommender system for sales team to find leads (**2.5x** efficiency increase)

Education

PhD Theoretical Physics | University of Nottingham

Oct 2019 – December 2023 Nottingham, UK

- **Thesis title: Quantum black holes**
- Designed and wrote large structures of C code to simulate the gravitational collapse of quantum matter
- Analysed and visualised simulation data using Python
- Published original research in high impact scientific journals (**3** papers)

MSc Theoretical and Mathematical Physics | Imperial College London

Sep 2018 – Sep 2019 London, UK

BSc Physics | University College London

Sep 2015 – June 2018 London, UK

Skills

Python: PyTorch, Tensorflow, Spark, Pandas, NLP, scikit-learn, Numpy, Matplotlib
Other: Git, AWS services, Airflow, Docker, C, OpenMP, SQL