SHARAN MAIYA

sm2783@cam.ac.uk

sharanm.uk ◇ github.com/maiush ◇ linkedin.com/in/sharanmaiya

EDUCATION

University of Cambridge *Oct 2023 - 2027 (expected)*
PhD AI for Environmental Risks
Supervised by Anna Korhonen, Ramit Debnath
Large Language Models for Decision-Making Under Risk in Climate Science

MRes Environmental Data Science
Supervised by Ramit Debnath, Laura Cimoli, Anna Korhonen
Thesis: Aligning Language Model Evaluators with Human Judgement

Imperial College London *Oct 2020 - Sep 2021*
MSc Statistics Merit
Supervised by Ioanna Papatsouma, D.K. Arvind
Thesis: A Novel Method of Tuning and Comparing Causal Discovery Algorithms on Real Data

The University of Edinburgh *Sep 2016 - Jun 2020*
BSc Computer Science and Mathematics First Class
Supervised by D.K. Arvind
Thesis: Investigating the Respiratory Rate Response to $PM_{2.5}$ Exposure in Asthmatic Adolescents

WORK + RESEARCH EXPERIENCE

Supervised Program for Alignment Research Oct 2023 - Jun 2024
Student Researcher

- Project 1: Towards a better understanding of sycophancy in large language models (LLM evals).
- Project 2: Investigating task breakdown in LLM's through model component clustering (mechanistic interpretability).
- Project 3: Contrast-pair clustering for CCS-style methods (concept-based interpretability).

Cambridge AI Safety Hub Oct 2023 - Dec 2023
Intro Fellowship Facilitator

- Teaching / guiding reading groups on literature in AI Safety.
- Topics covered both philosophical arguments and technical research.

The University of Edinburgh Sep 2021 - Jun 2023
Research Assistant

- Statistical methods and machine learning for a range of problems in air pollution epidemiology.
- Causal discovery algorithms and causal effect estimation.
- Debiased (targeted) machine learning for semi/non-parametric models.
- Advising undergraduates and masters students on a weekly basis.

TradingHub Jun 2020 - Aug 2020
Software Engineer Intern

DataGrasp Jan 2020 - Apr 2020
Freelance Data Scientist

Royal Bank of Scotland Jun 2019 - Aug 2019
Summer Intern

The University of Edinburgh Sep 2018 - Dec 2018
Undergraduate Researcher

PREPRINTS AND PUBLICATIONS

Walter Laurito, **Sharan Maiya**, Grégoire DHIMOÏLA, Owen Ho Wan Yeung, and Kaarel Hänni. "Cluster-Norm for Unsupervised Probing of Knowledge". *ICML Workshop on Mechanistic Interpretability* 2024, https://openreview.net/forum?id=kXRYju6Jtt.

D K Arvind and **S Maiya**. "Sensor data-driven analysis for identification of causal relationships between exposure to air pollution and respiratory rate in asthmatics". *arχiv* 2022, http://arxiv.org/abs/2301.06300.

D K Arvind, **S Maiya**, and P Sedeno. "Identifying causal relationships in time series data from a pair of wearable sensors". *IEEE 17th International Conference on Wearable and Implantable Body Sensor Networks* 2021, https://doi.org/10.1109/BSN51625.2021.9507030.

A Miller, D Miron, and **S Maiya**. "GraphDraw - A Tool for the Representation of Graphs Using Inherent Symmetry". In *Proceedings of The First International Conference on Symmetry*, 2018, https://doi.org/10.3390/proceedings2010086.