Dewi S. W. Gould

@dewi.s.w.gould@gmail.com, **List of Publications**

I am an **Astra AI Safety Fellow** working with Redwood Research. I have worked on explanatory faithfulness, scalable evaluations, inverse reinforcement learning, and computer vision. I have a strong background in mathematics and theoretical physics: I obtained a PhD in string theory in 2024 from the University of Oxford.

SELECTED RESEARCH

A Positive Case for Faithfulness: LLM Self-Explanations Help Predict Model Behavior
H Mayne, J.S. Kang, ***Dewi S.W. Gould****, et. al.*
- Submitted to ICML, 2026

PAC Apprenticeship Learning with Bayesian Active Inverse Reinforcement Learning
O Bajgar, ***Dewi S.W. Gould****, et. al.*
- Presented at **Reinforcement Learning Conference** 2025
- Presented at **NeurIPS Workshop** on Bayesian Decision Making, 2024

SKATE, a Scalable Tournament Eval: Weaker LLMs differential between stronger ones using verifiable challenges
Dewi S.W. Gould*, Bruno Mlodozeniec, Samuel Brown*
- In preparation

AirTrafficGen: Configurable Air Traffic Scenario Generation with Large Language Models
Dewi S.W. Gould *et. al.*
- Presented at **NeurIPS Workshop** on Language and World Models, 2025

EDUCATION

PHD in Mathematics
"Generalised Symmetries in String Theory Realizations of Quantum Field Theories"
University of Oxford (2020-2024)
- Research **summary**

MAst Applied Mathematics
University of Cambridge (2019-2020)
- *No numerical grades awarded due to COVID

MSci Theoretical Physics
Imperial College London (2015-2019)
- Highest rank final year student (97% final exam average)
- CMS Prize

SKILLS

Python, git, PyTorch.

EXPERIENCE

ASTRA AI SAFETY FELLOWSHIP, Constellation — January '26-Present, Berkeley, USA.

Working with **Redwood Research** on diffuse control and capability elicitation. Studying activation steering techniques, RLVR and distillation methods. Mentors: Julian Stastny and Alex Mallen.

SPAR AI SAFETY FELLOWSHIPS — January '25-Present, Glasgow, UK.
Two SPAR AI safety projects. The first, on scalable LLM evaluations mentored by Bruno Mlodozeniec (Uni. Cambridge) and Sam Brown (Independent), and the second on explanatory faithfulness mentored by Noah Siegel (GDM).

POSTDOCTORAL RESEARCHER, Alan Turing Institute. — June '24-December '25, London, UK.

Led research on autonomous methods for creating novel and controllable air traffic scenarios using Large Language Models. Helped engineering efforts in building a digital twin of UK airspace. Studied complexity prediction methods using graph neural networks. See outreach **talk.** Led by Prof. Richard Everson.