Theo Farrell

Manchester, UK | (+44)7596102384 | theo.farrell99@outlook.com
Website: https://theosdoor.github.io

Durham University MSci Natural Sciences student specialising in mechanistic interpretability of neural networks. Won L2 and L3 Natural Sciences Outstanding Achievement Awards. Founded Durham's AI Safety Initiative and secured an $6,500 grant from Coefficient Giving in 2025 for independent research. Recent L4 project research work at Durham University resulted in first-authorship of a paper for a NeurIPS 2025 workshop.

EDUCATION

Durham University 2022 – 2026
MSci Natural Sciences (Computer Science & Philosophy) Average grade: 1st (77%, 4.0 GPA)

- **Dissertation:** "Relational Composition and Sparse Autoencoders", supervised by Noura Al Moubayed.
- **Key Modules:** NLP, Deep Learning, Computer Vision, Reinforcement Learning.
- **Annual End-of-Year Averages:** Y1: 76%, Y2: 76.5%, Y3: 77%.
- **Top Module Rankings:** Natural Computing (1st/51), Computer Vision (2nd/81), Moral Theory (1st/135), Political Philosophy (2nd/138).

A-Levels Results 2020 – 2022
4 A-Levels (A*A*A*A*) including Further Maths, in addition to EPQ (A*).

EXPERIENCE

Founder and Lead Organiser 09/2023 – Ongoing
Durham AI Safety Initiative

- Organised recruitment and increased our weekly attendance from 5 to 20 participants.
- Facilitated BlueDot discussion groups to upskill members in AI alignment and policy knowledge.
- Led technical upskilling workshops in deep learning and NLP based on ARENA.
- Secured grant funding for group organising.
- Pathfinder Fellow since Dec 2024 - received training and mentorship for AI Safety organising after a competitive application process.

University of Calgary 06/2024 – 09/2024
MITACS Global Research Intern (supervised by Aditya Nittala)

- Co-author on FabFoam (2025): used Arduino and Java to create and program soft electrotactile devices that provide haptic feedback.
- Designed a VR application using A-Frame (HTML and JavaScript) on the Meta Quest 3 for a paper at ACM UIST'24 (acknowledged, https://doi.org/10.1145/3654777.3676448).

PUBLICATIONS [Scholar]

- Farrell, Theo, Patrick Leask, and Noura Al Moubayed. "Order by Scale: Relative-Magnitude Relational Composition in Attention-Only Transformers." Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025. Link: https://openreview.net/forum?id=vWRVzNtk7W
- Kempermann, Manon, et al. "Challenges of Evaluating LLM Safety for User Welfare." Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI'26), 2026. Link (preprint): https://arxiv.org/abs/2512.10687

Reviewing Activities: NeurIPS 2025 Mechanistic Interpretability Workshop, NeurIPS 2025 ResponsibleFM Workshop.

GRANTS AND AWARDS

- **$4,500 from Pathfinder** (10/2025): 1 year of group expenses for Durham AI Safety Initiative.
- **$6,500 from Coefficient Giving (fmr. Open Philanthropy)** (07/2025): 4 months of independent mechanistic interpretability research.

- **$8,000 from Coefficient Giving (fmr. Open Philanthropy)** (03/2025): 1 year of group expenses for Durham AI Safety Initiative.
- **L3 Natural Sciences Outstanding Achievement** (07/2025): Averaged over 75% in 3rd year.
- **L2 Natural Sciences Outstanding Achievement** (07/2024): Averaged over 75% in 2nd year.

CERTIFICATIONS

- **AGI Strategy** (BlueDot Impact, 2026): 30-hour course covering technical AI trends and future capabilities, threat modelling via kill chain analysis, and defence-in-depth frameworks for designing protective interventions. [Certificate]

VOLUNTEERING

Community Outreach
County Durham

- Supported local community initiatives, including County Durham Foodbank (warehouse support, sorting donations and packing food parcels) and Sherburn House Charity (resident engagement through companionship activities such as conversation, games, and walks).

A-Level Maths Tutor at local state schools
County Durham

- I regularly volunteer as a tutor for sixth-form students at local state schools, most recently delivering A-level Maths tuition at Durham Johnston Comprehensive School through a ten-week programme of weekly one-hour sessions.
- Designed lesson plans according to their individual needs and communicated complex ideas in an understandable and accessible manner.

TECHNICAL SKILLS

Programming (Fluent)	Python, C#, Javascript, HTML.
Programming (Familiar)	C, C++, Java, Haskell, Fortran, Ruby.
Tools	PyTorch, TransformerLens, SAELens, TensorFlow, scikit-learn, Hugging Face, NumPy, Weights & Biases, OpenCV, Seaborn, matplotlib, pandas, Git, uv, LaTeX, AR/VR, Unity, Arduino.