Theo Farrell

theo.farrell99@outlook.com | https://theosdoor.github.io

Last updated: 27/11/2026

AI safety researcher and PhD student at Durham University, working on control and monitoring of LLMs. First author of two mechanistic interpretability papers (NeurIPS 2025 and ICML 2026 workshops), and a paper under review evaluating model organisms of misalignment completed as a Research Fellow at LASR Labs. Founded Durham's AI Safety Initiative and secured $19,000 in research and organising grants.

EDUCATION

Durham University 2026 – Present
PhD Computer Science (supervised by Noura Al Moubayed) Control and monitoring of LLMs

Durham University 2022 – 2026
MSci Natural Sciences (Computer Science & Philosophy) First Class Honours (Class I, 79%)

- **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%, Y4: 82%.
- **Top Module Rankings:** Natural Computing (1st/51), Computer Vision (2nd/81), Moral Theory (1st/135), Political Philosophy (2nd/138).

EXPERIENCE

LASR Labs Summer 2026
Research Fellow, London (mentored by Satvik Golechha)

- Designed evaluations measuring how reliably model organisms exhibit an implanted behaviour (paper under review).

Founder and Lead Organiser 09/2023 – 02/2026
Durham AI Safety Initiative

- 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.
- Pathfinder Fellow from 12/2024: 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)

- Built soft electrotactile haptic devices; contributed to a paper at ACM UIST'24.

RELEVANT PUBLICATIONS [Scholar]

- Farrell, Theo, Patrick Leask, and Noura Al Moubayed. "Sparse Autoencoders Can Learn Graded Latents for Relational Composition." Mechanistic Interpretability Workshop at ICML 2026. Link: https://openreview.net/forum?id=ltQ6XAduTF
- 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. "Safe for whom? Rethinking how we evaluate the safety of LLMs for real users." Second Conference of the International Association for Safe and Ethical Artificial Intelligence (IASEAI'26), 2026. Link: https://ojs.aaai.org/index.php/IASEAI/article/view/43033

Reviewing Activities: ICML 2026 Mechanistic Interpretability Workshop, 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** (07/2025): 4 months of independent mechanistic interpretability research.
- **$8,000 from Coefficient Giving** (03/2025): 1 year of group expenses for Durham AI Safety Initiative.
- **L4 Natural Sciences Outstanding Achievement** (07/2026): Averaged over 75% in 4th year.
- **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.

RELEVANT TECHNICAL SKILLS

Programming	Python, C#, JavaScript, HTML.
Tools	PyTorch, TransformerLens, SAELens, Hugging Face, NumPy, scikit-learn, Weights & Biases, pandas, Git, uv, LaTeX.