Lucy Farnik

+44 7399 156410
London, UK

LinkedIn
lucyfarnik@gmail.com

Profile

I am a PhD student passionate about post-training, continual learning, and model robustness. I've published multiple papers at ICLR/ICML in areas from RL to interpretability and also collaborated with researchers from Oxford, Berkeley, and Google DeepMind. My previous background is in software engineering – I started coding at age 7 and became a senior developer at age 18.

Publications

Jacobian Sparse Autoencoders: Sparsify Computations, Not Just Activations
***L. Farnik**, T. Lawson, C. Houghton, L. Aitchison* *ICML 2025*
- Created a new method for discovering small interpretable units of computation in neural nets

Residual Stream Analysis with Multi-Layer SAEs
*T. Lawson, **L. Farnik**, C. Houghton, L. Aitchison* *ICLR 2025*
- Studied information flow in LLMs by training a single SAE on all layers

Sparse Autoencoders Can Interpret Randomly Initialized Transformers
*T. Heap, T. Lawson, **L. Farnik**, L. Aitchison* *Under review at NeurIPS 2025*
- Demonstrated that standard methodology around SAEs can be misleading

STARC: A General Framework For Quantifying Differences Between Reward Functions
*J. Skalse, **L. Farnik**, S. Motwani, E. Jenner, A. Gleave, A. Abate* *ICLR 2024*
- Created a set of reward distance metrics which exceeded the previous state of the art

Inducing Human-like Biases in Moral Reasoning Language Models
*A Meek, A Karpov, S. Cho, R. Koopmanschap, **L. Farnik**, B. Cirstea* *NeurIPS UniReps workshop 2024*
- Fine-tuned LLMs on fMRI data in order to induce more human-like behavior

Engineering Experience

Senior Full Stack Developer August 2019 – June 2023
Longitude 103 *Remote*
- Became a senior developer at age 18
- Adapted to leading multiple B2B platforms in parallel, including building several products on my own, working across front end, back end, DevOps, iOS, Android, and UX
- The primary project I led was bringing in ∼$300,000 in annual revenue – nearly 10x more than when I joined the company, largely due to the new features and UX improvements I built

Research Experience

PhD Student September 2023 – Present
University of Bristol *Bristol, UK (remotely from London)*
- Researching LLMs and LLM agents under Dr. Laurence Aitchison
- The 2023-24 academic year was a taught qualifications year which I completed with first-class honours

MATS Scholar (Neel Nanda stream) November 2023 – July 2024
ML Alignment & Theory Scholars *Berkeley, CA*

- Exploring SAE-based circuit analysis and static analysis under Neel Nanda (Google DeepMind) and Arthur Conmy (Google DeepMind)

Research Lead & Cofounder August 2023 – January 2024
Bristol AI Safety Centre (BASC) *Bristol, UK*

- Co-leading a small research lab focused on AI security; funded by Lightspeed Grants

Forecasting paradigm shifts in AI July 2023 – September 2023
Epoch AI and the Forecasting Research Institute (FRI) *Remote*

- Statistical modeling for predicting the frequency of large paradigm shifts in ML

Research Intern July 2023 – September 2023
University of Bristol *Bristol, UK*

- EPSRC-funded internship researching representations of grammar in LLMs

Researcher May 2023 – June 2023
Alignment Research Engineer Accelerator (ARENA) *London, UK*

- High-intensity ML research program in the London SERI MATS office

Researcher February 2023 – June 2023
AI Safety Camp *Remote*

- Fine-tuning LLMs on a moral reasoning dataset using fMRI of humans performing the same tasks

Grants

Long-Term Future Fund April 2024

- $41,000 grant for my research on sparse circuit discovery during MATS

Lightspeed Grants August 2023

- $60,000 award to support my research on AI robustness

Lightspeed Grants August 2023

- $12,000 grant for BASC, a small research organization I co-founded

Education

University of Bristol 2023 – 2027
PhD Interactive Artificial Intelligence *Bristol, UK*

- My PhD is focused on deep language modeling; I'm supervised by Dr. Laurence Aitchison
- 2023–24 was a qualification/coursework year which I completed with first-class honours

University of Bristol 2020 – 2023
BEng Computer Science with Innovation *Bristol, UK*

- First-class honours
- Relevant units and marks: Data-Driven Computer Science (97), Artificial Intelligence (92), Machine Learning (85), Computer Architecture (96), Image Processing and Computer Vision (86)