

Patrick Leask - Curriculum Vitae

Personal Information

Email patrickaaleask@gmail.com
LessWrong <https://www.lesswrong.com/users/patrick-leask>

Education

2023 - 2026 **Machine Learning PhD**
Durham University - Supervised by Noura Al Moubayed and David Budgen

2021 - 2022 **Theoretical Computer Science MSc (Distinction)**
University of Liverpool - Supervised by Xiaowei Huang

2014 - 2018 **Statistical Science (International Programme) MSci. 2.1 (Hons)**
University College London, studied abroad at Nanyang Technological University - Supervised by Franz Kiraly

Selected Publications [\[Scholar\]](#)

- Inference Time Decomposition of Activations: A Scalable Approach to Interpreting Large Language Models [\[Arxiv\]](#), first author, ICML 2025 poster
- Sparse Autoencoders Do Not Find Canonical Units of Analysis [\[Arxiv\]](#), joint first author, ICLR 2025 Poster
- Probing by Analogy: Decomposing Probes into Activations for Better Interpretability and Inter-Model Generalization [\[OpenReview\]](#), first author, NeurIPS 2025 Mechanistic Interpretability Workshop Spotlight
- BatchTopK SAEs [\[Alignment Forum\]](#), co-author, NeurIPS SciForDL'24

Reviewing Activities: ICML 2026, NeurIPS 2025 Mechanistic Interpretability Workshop, COLM 2025 Socially Responsible Language Modelling Research, COLM 2025 Workshop on the Application of LLM Explainability to Reasoning and Planning, NeurIPS 2024 Workshop on Scientific Methods for Understanding Deep Learning

Work Experience

08/2026-Ongoing **Founder**
AI Assurance Lab (website TBD)

- Researching how and why to audit AI labs.

04/2026-Ongoing **Co-Founder (Part-Time)**
CEasy <https://ceasy.onrender.com/>

- Developing software for recreational watercraft regulatory compliance and documentation

01/2026-06/2026 **Astra Research Fellow**
<https://www.constellation.org/programs/astra-fellowship>

- Funded research fellowship in Berkeley
- Researched AI control under the mentorship of Charlie Griffin of UK AISI

03/2024-11/2025 **Scholar**
MATS <https://matsprogram.org/>

- Fully funded research fellowship in Berkeley
- Research sparse dictionary learning for mechanistic interpretability
- Mentored by Neel Nanda of Google Deepmind

09/2022-12/2022 **AI Technical Alignment Research Intern**
Aligned AI <https://www.aligned-ai.com/>

- Conducted research into concept extrapolation for deep learning models in imaging and reinforcement learning settings.

02/2021- 06/2022

Back-End Engineer

MachineMax <http://www.machinemax.com>

- Developed event-driven Go & Python microservices deployed on GCP to manage and analyse sensors in heavy industry.
- Led features including a redesigned ABAC system, worked on performance issues with services and Postgres, and contributed to improving monitoring and observability.

09/2018 - 06/2020

Back-End Engineer

Thought Machine <https://www.thoughtmachine.net>

- Worked on smart contracts for retail banking in Python and Go.
- Involved in recruiting new engineers and training them in Go programming.