

Matthew Southerby

Research Physicist

Physicist with strong analytical and quantitative foundations seeking a role in AI safety research. Particular interest in mechanistic interpretability, model personas, and the emerging questions around digital minds. Independent research includes extracting and causally intervening on concept manifolds in transformer residual streams, and probing mimicry via trait vectors in multi-turn conversations (matthewsoutherby.com). Experienced in driving research across multi-disciplinary groups, with a commitment to continuous professional growth and the ability to quickly adapt to new domains.

Independent Research

- Analysed the geometry of concept families (days, months, compass directions) in the residual streams of a 13B model, showing structured low-dimensional manifolds with interpretable geometric properties across layers.
- Demonstrated causal intervention on a 2D concept manifold: rotating steering vectors in PCA space produces smooth sinusoidal changes in output probabilities, suggesting the model causally reads from these subspaces during inference.
- Investigated tonal mimicry and sycophancy in multi-turn conversations by extracting trait vectors and measuring how assistant activations track escalating user tone. Found evidence that sycophancy is a multilayer concept rather than a single steerable direction.

Professional Experience

Research Physicist [University College London Hospital](#) **London, UK** *September 2024 - present*

- Automated patient treatment workflows for various treatment modalities using software development and LLM-based tooling, including retrieval-augmented generation and Model Context Protocol pipelines.
- Automated clinical quality assurance tests using AI classification methods, significantly reducing effective treatment time.
- Drove collaboration within a multi-disciplinary research group and translated research into the clinic.

Postdoctoral Research Associate [Lancaster University](#) **Lancaster, UK** *January 2024 - August 2024*

- Contributed to cutting-edge research on particle accelerators and their applications.
- Collaborated with researchers all over the world on projects, such as optimizing accelerator cavity design with CERN.
- Led research experiments with junior scientists and non-technical stakeholders such as collaborating with Gucci on an environmentally friendly method for tanning of leather.

Education

PhD — Accelerator Physics [Lancaster University](#) **Lancaster, UK** *2020–2024*

Thesis Title: "Beam Dynamics Studies and RF Cavity Design for Proton Radiotherapy Machines"

Combined analytical methods with computational modelling to push advancements in particle accelerator physics.

MSci — Physics — First Class [University of Bristol](#) **Bristol, UK** *2016–2020*

Final Year Project: "Tracking at Lepton Colliders with Machine Learning" — First Class

Gained experience in machine learning methods applied to particle tracking.

Skills and Interests

- **Technical Skills:** Python, MATLAB, C#, C++; experienced in data collection, cleaning, processing, and visualisation; proficient in developing and automating research workflows using LLM-based APIs and Model Context Protocol; hands-on experience with transformer internals including contrastive activation extraction, steering vectors, PCA of residual streams, and causal intervention via activation patching.
- **Analytical Skills:** Statistical analysis, machine learning, computational modelling, mechanistic interpretability (concept geometry, trait vector extraction, sycophancy probing in multi-turn settings), data-driven inference, optimisation methods, and visualisation of high-dimensional activation spaces.
- **Soft Skills:** Experienced in leading multidisciplinary research projects, communicating technical concepts clearly, managing stakeholder expectations, and mentoring junior colleagues.

Journal Publications

- Southerby, M, et al. "Multi-Objective Optimization and Pareto Front Visualisation Techniques applied to Normal Conducting RF Accelerating Structures." *Phys. Rev. Accel. Beams* (June 2022).
- Southerby, M & Apsimon. R. "Beam dynamics framework incorporating acceleration to define the minimum aperture in two focusing schemes for proton radiotherapy linac." *Phys. Rev. Accel. Beams* (June 2024).
- Dean, C.B.C., Pérez-Lara, M.L., Horton, E., Southerby, M, et al. "SDE-based Monte Carlo dose calculation for proton therapy validated against Geant4." *arXiv:2511.03115* (2025).