

Professional Experience

AI Safety Research Fellow, Athena programme, November 2025 – Present

Trained conditional diffusion models to investigate compositional generalization.
Exploring interpretability of training dynamics for emergent structure.
Mentored by Jesse Hoogland (Timaeus) - publication in progress.

Data Scientist, VodafoneThree, Mar 2025 – Present

Built large-scale ML pipelines improving campaign performance by 30%.
Applied LLM-based analysis of customer calls to identify revenue increase opportunities.
Clearly communicated technical insights and methodological limitations to non-technical stakeholders.

Independent Researcher, July – September 2025

Applied Sparse Autoencoders to biological neural data to extract latent features.
Produced a NeurIPS workshop paper and tutorial.

AI Safety Research Fellow, SPAR, Feb - June 2025

Mechanistic interpretability experiments in small neural networks.
Performed targeted ablation studies and neuron weights and activations visualizations to study feature superposition.
Mentored by Stefan Heimersheim (FARAI) – NeurIPS workshop paper.

PhD Researcher, University College London, Sep 2019 – Dec 2024

Independently led research on internal representations in biological neural networks.
Developed end-to-end analysis and modelling methods on high-dimensional neural population data, including Bayesian inference and dimensionality reduction.

Machine Learning Engineer Intern, Open Climate Fix, Aug – Oct 2022

Trained and evaluated ML forecasting models for solar energy production.

Education

PhD Systems and Computational Neuroscience, UCL & QMUL (UK), 2019 – 2024

Funded by LIDo Doctoral Training programme (5% acceptance rate)

MSc Neuroscience (Distinction), UCL (UK), 2018 - 2019

BSc Biomedical Sciences (First Class Honours), UAB (Spain), 2014 – 2018

Selected Publications in Mechanistic Interpretability and Deep Learning (* denotes equal contribution)

- **Molas-Medina**. Training dynamics and phase transitions in compositional generalization of diffusion models (In prep).
- Bhagat*, **Molas-Medina***, Giglemani, Heimersheim. Compressed Computation is (probably) not Computation in Superposition. *NeurIPS, Mechanistic Interpretability workshop paper*, 2025.
- Bhagat, Pouget, **Molas-Medina**. A pipeline for interpretable neural latent discovery. *NeurIPS, Data on the Brain & Mind findings workshop paper*, 2025.
- Pouget, Bhagat, **Molas-Medina**. NLDISCO: A pipeline for interpretable neural latent discovery. *NeurIPS, Data on the Brain & Mind findings workshop tutorial*, 2025.

Additional Research Activities

- Open Philanthropy technical AI safety RFP Research Grant (Final Round, 2025)
- Peer reviewer: NeurIPS 2025 workshop UniReps: Unifying Representations in Neural Models
- ARENA (AI alignment Research Engineer Accelerator) course (2025)
- Machine Learning Summer School (Stellenbosch University, South Africa, 2023)

Machine Learning & AI

Mechanistic interpretability
Deep learning
Supervised learning
Unsupervised learning

Engineering

Python (incl. PyTorch, sklearn, numpy, pandas)
SQL
Data/ML pipelines
Version control (Git)

Data analysis

Statistics
Data processing
Data exploration
Predictive modeling
Experimental design

Additional skills

Biomedical sciences (molecular & cell biology, genetics, physiology)
Independent research
Group research
Cross-disciplinary