Edward Stevinson

edward.stevinson [at] hotmail [dot] com | *github.com/Stevinson* | *stevinson.github.io*

Profile

Machine learning researcher and engineer with nearly a decade of experience, currently pursuing a PhD in mechanistic interpretability focusing on the effect of superposition on adversarial examples. My work spans from founding a company to roles at startups and established firms, giving me exposure across the entire company life cycle, helping to develop a diverse range of technical skills and familiarity with tools across all stages of ML development. Above all, I am motivated by theoretically challenging and novel projects, and implementing research as part of a team.

Professional Experience

Safe Intelligence - Research Scientist
Mar 2023 -

- Lead developer of our neural network verification and certified training toolkit. The role focused on performing and productionising research that ensures neural networks behave as expected under input perturbations and adversarial attacks.

Iris - Research Scientist Intern
Jul 2022 - Oct 2022

- Research focused on leveraging knowledge graphs to update word embeddings in the scientific domain concluding with presenting our paper at an AACL workshop on information extraction.

Napo - Senior Software Engineer
Apr 2022 - Sep 2022

- Contract python backend developer for the fast growing pet insurance startup. As a member of the growth team, grew the customer base from 1k to 10k over six months.

Upside Ltd - Lead Data Scientist and Co-Founder
Nov 2020 - Feb 2022

- A workflow analytics platform that helps empower teams by surfacing key metrics and highlighting trends. My work included creating the Python analytics engine that underpin the platform, the data science pipelines that consume data from the engine, and the models used by the platform.

Eigen Technologies - Solutions Engineer
Jun 2019 – Nov 2020

- A research-led artificial intelligence company that automates the extraction and classification of information from documents, specialising in small training sets.
- Developed a tool for sectioning legal documents and producing a hierarchical representation.
- Built a system to manage the company's ML lifecycle to make model development less costly, achieved by using a custom setup comprised of Pachyderm, MLFlow, Cookiecutter and AWS.

IBM - Software Developer
Jun 2017 – Jan 2019

- Java developer for IBM's i2 brand that produces investigative analysis software for law enforcement and intelligence agencies worldwide.

Work experience:
3Space (2015), UBS AG (2012), Royal Bank of Scotland plc (2010), Southampton University (2009), Atkins Oil & Gas (2008)

Education

Imperial College London

2022- Imperial-X PhD scholarship in Safe AI researching mechanistic interpretability and robustness under the supervision of Professor Tolga Birdal and Professor Alessio Lomuscio.

University of Edinburgh.

Artificial Intelligence (MSc) - 1. Distinction

2015-2016 Machine learning specialism with a focus on multi-agent systems and game theory.

Trinity College, Oxford.

Engineering, Economics and Management (MEng) - 2.1

2011-2015 Specialised in machine learning, robotics and algorithmic game theory.

Dulwich College, London.

2006-2011 A-levels: 4 A*; GCSE: 11 A*, 1 A

Technical Skills

- 7+ years professional coding experience.
- 5+ years professional machine learning experience, predominantly PyTorch.
- Experience deploying ML solutions using Docker, Airflow and Dagster.
- Utilise data science best practices and industry-standard tools such as Weights & Biases, Cookiecutter, JupyterLab, MLFlow, Kafka and Pachyderm.
- Maintain strict development best practices surrounding testing, versioning, and Agile.
- Strong mathematical background with modules covering reinforcement learning, supervised & unsupervised probabilistic modelling, neural networks, sampling and graph theory.
- A keen interest in novel projects. From founding Upside, back through IBM patent submission groups, to the 2010 HSBC Young Enterprise Scheme - all have involved novel products.
- Strong writing skills built upon in my management studies, dissertations and grant proposals.

Publications

Stevinson, E, and Lomuscio, A, 2024, Reducing Return Volatility in Neural Network-Based Asset Allocation via Formal Verification and Certified Training, In Proceedings of ICAIF. ACM, New York, NY.

Hoelscher-Obermaler, J, and **Stevinson, E**, 2022, Leveraging knowledge graphs to update scientific word embeddings using latent semantic imputation, Workshop on Information Extraction from Scientific Publications at AACL-IJCNLP 2022

Teaching

- Deep Learning. T.A., Imperial College London, 2023.
- Natural language processing. T.A., Imperial College London, 2023.
- Demystifying machine learning. T.A., Imperial College London, 2023.
- Codebar programming mentor, 2020-2022.

Projects

2024 **ViT-Prisma**

Open source contributor to the Prisma toolkit - a mechanistic interpretability library for multimodal models.

Reprogramming AI Models Hackathon - Apart Research

Explored how LLMs respond to adversarial attacks at the feature level. Working with Goodfire's SDK, our second place entry investigated whether latent activations could help identify harmful prompts.

2023 **AI Safety Camp**

Drafting high-risk policy proposals for the EU Safety Act as part of AISC8.

2020 **Forfit**

Developed the backend for a challenge-based fitness app using Django.

2019 **Credible Choice**

Developed the API in Go for a real-time, accurate register of Brexit views to allow UK citizens to express their opinions and raise money for charity.

2018 **Towards Fair AI: tackling bias in predictive probabilistic models**

Essay submission for GeneralAI's 'mitigating the risks of the AI race' competition.

2017 **Unpolarise**

Developed an app to reduce the echo chamber effect by summarising the news you consume and challenging your newsfeed bias with recommendations outside your conversational sphere.

2016 **Utilising Policy Types to Achieve Effective Ad Hoc Coordination**

Research dissertation that involved building an autonomous agent able to maximise payoff in multi agent systems with no mechanisms for prior coordination.