# Callum McDougall

+44 7985 552727 | cal.s.mcdougall@gmail.com | GitHub callummcdougall | Personal website perfectlynormal.co.uk

## EDUCATION

**University of Cambridge** **Cambridge, UK**
*Master of Mathematics* *2018 – 2022*
*Year 4: Distinction, Year 3: Class I [rank 11/222], Year 2: Class I [honorary], Year 1: Class I [rank 38/232]*

- **Awards:** Dr J.A.J. Whelan Prize in Mathematics, Christ's College Academic Scholarship (1st, 2nd & 3rd years)
- **Courses:** Statistics, Optimisation, Applied Probability, Automata & Formal Languages, Logic & Set Theory

**Westminster School** **London, UK**
*STEP: S, S ("Outstanding") grades in papers II and III (top 15 in the country)* *2013 – 2018*
*A Levels: Maths (A*), Further Maths (A*), Economics (A*), Physics (D1, equivalent to high A*)*

- UKMT: Certificate of Distinction in 2nd round of British Mathematical Olympiad (top 25 of 30,000 students)
- Senior Cheyne Prize for Mathematical Excellence (awarded to one senior student annually)

## EXPERIENCE

**Alignment Research Engineer Accelerator (ARENA)** **London, UK**
*Director* *Sep 2022 – Jan 2022, May 2023 - Jun 2023*

- Founded & ran the first and second iterations of ARENA (with Kathryn O'Rourke as head of operations)
- Acted as head TA and curriculum designer for the majority of the program
- Main roles included:
    - Assisting participants with conceptual and engineering questions
    - Structuring the micro and macro-level details of the curriculum
    - Giving several talks and organizing group discussions
    - Providing feedback and advice to groups while they were working on capstone projects
- Created materials for people studying the program virtually, including a monthly series of mechanistic interpretability challenges
- Designed our public-facing website here, and the website for hosting the curriculum here (which now has over 2000 unique visitors)

**SERI MATS** **London, UK (virtual)**
*Scholar* *Jun 2023 – Aug 2023*

- Accepted into Neel Nanda's Mechanistic Interpretability SERI MATS stream with Arthur Conmy, studying negative attention heads in GPT2-Small
- Developed a form of ablation which demonstrates that a particular mechanism (copy suppression) explains the majority of head L10H7's behaviour in GPT2-Small, and reoccurs as a motif in larger models
- Wrote a paper on which I will be first author (draft available here)

**Winter ML Bootcamp** **London, UK**
*Head TA (Boston)* *Jan 2022 – Feb 2022*

- Invited to the Winter ML Bootcamp run by the MIT AI Alignment group, as head TA in Boston
- Received very positive feedback from participants
    - All five participants who left comments in the TA feedback forms specifically mentioned that I had given helpful talks or explanations
- Expanded parts of the curriculum by combining my writing on induction heads with Neel Nanda's material on TransformerLens and mechanistic interpretability
- Stayed on for two weeks afterwards, to help run the mechanistic interpretability workshop & project week

**MLAB2** **London, UK**
*Participant* *Aug 2022 – Sep 2022*

- Participated in the second Machine Learning Alignment Bootcamp
- Activities included assembling and training transformers, and doing transformer interpretability exercises based on the material in A Mathematical Framework for Transformer Circuits

**AI Safety Camp** **Cambridge, UK**
*Participant* *Jan 2022 – present*

- Working in a team of three, researching selection for modularity in evolved systems & modern deep learning
- Have given several summary talks on our team's progress
- Responsible for two LessWrong posts (links here and here)
    - The first one sketches out different theories for modularity in the biological literature
    - The second one summarises our team's approach and current results

**SERI MATS** **Cambridge, UK (virtual)**
*Scholar* *Oct 2021 – Dec 2021*

- Attended weekly discussions with Evan Hubinger to discuss his research agenda, and his views on alignment
- Wrote up a distillation of John Wentworth's *Natural Abstraction Hypothesis*, linking it to other topics such as interpretability research and neuroscience (link here)
    - This distillation was featured for a time as part of the AGISF 201 course

**Cambridge Existential Risks Initiative** **Cambridge, UK**
*Committee member* *Sep 2021 – Sep 2022*

- Designed syllabus for & organised an introductory course on existential risks
- Organising CERI activities such as the AGI Safety Fundamentals Programme
- Co-leading a reading group on *Human Compatible: Artificial Intelligence and the Problem of Control*

**Jane Street Capital** **London, UK**
*Quantitative Trading Intern* *Jun 2021 – Aug 2021*

- Participated in team-based mock trading sessions, requiring delegation and coordination
- Built a volatility model for corporate bonds using Gradient Boosted Trees
- Developed methods for interactive visualisation of the GBT result in Python
- Received a full-time offer as a trader

**Effective Altruism, Cambridge** **Cambridge, UK**
*Facilitator* *Jan 2020 – Mar 2020*

- Facilitated the Effective Altruism Introductory Fellowship for two groups of Cambridge residents
- Helped to promote discussion of important topics, and helped fellows engage with ideas
- One fellow is now a member of the EA Cambridge Extended Committee

## SKILLS, ACTIVITIES & INTERESTS

**Technical Skills:** Python, Excel, VBA, SQL, LaTeX, JavaScript, HTML
**Certifications & Training**: Guinness World Record holder
**Activities:** member of Cambridge Existential Risks Initiative, Effective Altruism
**Interests:** chess, rock-climbing, mathematical art