Doctoral student pursuing an academic career understanding risks from advanced artificial intelligence. Academic website: https://far.in.net.

§ Education

Doctor of Philosophy in Computer Science Oct 2024 (expected)–2027 (expected)
University of Oxford
Understanding emergent agency in learned AI systems. Supervised by Professor Alessandro Abate.

Master of Computer Science (with Distinction) Part-time 2019–2022
The University of Melbourne Coursework average **98.8%**, thesis **95.5%**
ETH Zürich (semester exchange, 2020) GPA **5.9** / 6.0
Advanced coursework in computation and learning. Thesis in deep learning theory (5) supervised by Daniel Murfet, leading to two sole-author conference papers (2, under review; 3, NeurIPS 2023). Founded a reading group on AI safety.

Awards: *Dean's Honours List* (top 5% marks in Faculty of Engineering and IT). *Top thesis mark* since the degree was first conferred in 2021. Thesis mark in the 95%+ category, reserved for theses described as follows: *"Truly outstanding in every way. In an entire academic career such a student may be encountered only once or twice. The student would be welcome as a PhD candidate in the School and would be expected to succeed with a hands-off supervision style."*

Bachelor of Science (Computing and Software Systems) 2014–2016
The University of Melbourne Average **93%**
Major in computer science and software engineering. Electives mainly in mathematics and physics.

Awards: *Dean's Honours List I, II, III* (top 1% marks in Faculty of Science in first, second, and third year). *Australian Computing Society Bachelor of Science Student Award* (top marks in third-year computer science coursework). *Australian Artificial Intelligence Institute Prize* (top marks in AI coursework). Top marks in many other computer science classes.

Victorian Certificate of Education (secondary school) 2013
Mount Lilydale Mercy College National percentile **99.8$^{\text{th}}$**
Maths/Science Prefect (elected by peers). Initiated/presented mathematics exam revision lecture.

Awards: *Dux* (valedictorian). *Victorian Premier's Award (Physics)* (top 3 physics students, state). *Australian Student Prize* (top 500 students, national). *Australian Defence Force Long Tan Leadership and Teamwork Award* (recognising leadership and contribution to school community).

§ Research Experience

Research Assistant (AI alignment & reward hacking) Sep 2023–present
Krueger AI Safety Lab & Computational and Biological Learning Lab, University of Cambridge
Working on understanding and mitigating goal misgeneralisation in deep reinforcement learning.

Research Associate and Research Lead Aug 2023–present
Timaeus
Working on understanding the emergence of in-context learning in transformers and other projects.

Research Assistant (Human-agent interaction) Jan 2023–Jul 2023
School of Computing and Information Systems, the University of Melbourne
Contributed to ongoing explainable AI project, evaluating human understanding of automated decision-making systems. Automated the creation of dynamic surveys with thousands of variants.

Virtual Research Intern Jun 2021–Oct 2021
Center for Human-compatible AI, University of California (Berkeley)
Project work leading to a paper on reward learning theory (4, ICML 2023). Initiated a virtual mini-conference for interns to share presentations about their projects.

See also Master of Computer Science (Master's research project). Part-time Feb 2021–Oct 2022

§ Publications

Machine Learning

(1) Jesse Hoogland,[(=)] George Wang,[(=)] **Matthew Farrugia-Roberts,** Liam Carroll, Susan Wei, Daniel Murfet, 2024, "The developmental landscape of in-context learning". Preprint arXiv:2402.02364. Under review.

(2) **Matthew Farrugia-Roberts,** 2023, "Proximity to losslessly compressible parameters". Preprint arXiv:2306.02834. Under review.

(3) **Matthew Farrugia-Roberts,** 2023, "Functional equivalence and path connectivity of reducible hyperbolic tangent networks". Preprint arXiv:2305.05089. Conference paper, **NeurIPS 2023.**

(4) Joar Skalse,[(=)] **Matthew Farrugia-Roberts,**[(=)] Alessandro Abate, Stuart Russell, and Adam Gleave, 2023, "Invariance in policy optimisation and partial identifiability in reward learning." Preprint arXiv:2203.07475. Conference paper, **ICML 2023.**

(5) **Matthew Farrugia-Roberts,** 2022, *Structural Degeneracy in Neural Networks*, Master's thesis, School of Computing and Information Systems, the University of Melbourne. Available online.

Computer Science Education

(6) **Matthew Farrugia-Roberts**, Bryn Jeffries, and Harald Søndergaard, 2022, "Teaching simple constructive proofs with Haskell programs." Extended abstract presented at TFPIE 2022, journal paper published in EPTCS. doi:10.4204/EPTCS.363.4.

(7) **Matthew Farrugia-Roberts**, Bryn Jeffries, and Harald Søndergaard, 2022, "Programming to learn: Logic and computation from a programming perspective." Conference paper presented at ITiCSE 2022, published by ACM. doi:10.1145/3502718.3524814.

§ Teaching Experience

Teaching Assistant and Guest Lecturer 2021, 2023–2024
Centre for AI and Digital Ethics, the University of Melbourne
Facilitating classes in the ethics and governance of AI for technical graduate students. Guest lecture on ethics and the future of intelligence (recording available online).

Sessional Subject Coordinator and Lecturer Jan 2018–Jul 2018
School of Computing and Information Systems, the University of Melbourne
Co-coordinated/lectured a summer intensive class on introductory programming (150 students). Co-coordinated a semester-long class on algorithms and data structures (400 students).

Head Teaching Assistant 2017–2021
School of Computing and Information Systems, the University of Melbourne
Designed coursework/assessment for classes on algorithmics, theoretical computer science, and artificial intelligence (300–600 students/class). Coordinated tutor teams to assess students fairly. Pioneered digital teaching/assessment methods leading to two CS education publications (6; 7).

Awards: *School of Engineering Tutor Community Excellence Award (finalist). School of Engineering Most Innovative Academic (finalist). Head Tutor Special Commendation Award.*

Teaching Assistant 2016–2021
School of Computing and Information Systems, the University of Melbourne
Taught above-listed classes plus classes in programming, operating systems, networks, and security.

Awards: *School of Computing and Information Systems Excellence in Tutoring Award, 2016.*

International Volunteer English Teacher Feb 2016
The Green Lion, Sri Lanka
Volunteer 4-week tour in Sri Lanka. Contributed to community teaching programs.

Volunteer Residential Mentor Jan/Jul, 2016–2018
Strengthening Engagement & Achievement in Mathematics & Science, the University of Melbourne
Volunteered to provide academic and pastoral support to under-represented mathematics and science secondary students during holiday study support retreats at the University of Melbourne.

Mathematics and Physics Presenter Dec 2014–Nov 2016
ATAR Notes (Australian student support community)
Created accessible lectures, subject notes, and webinars for secondary students.

Private Tutor and School-based Tutor Jan 2014–Jun 2017
Private & Mount Lilydale Mercy College & Scotch College (indigenous student support program)
Mathematics and study-skills support for secondary students of diverse backgrounds. Volunteered to organise annual final exam revision classes, including delivering mathematics and physics classes and recruiting other high-achieving alumni to deliver classes on other topics.

§ Technical Proficiencies

Programming: Python (including NumPy, SciPy, matplotlib, PyTorch, PyTorch/XLA, einops, JAX); C; Haskell; JavaScript (including React).

Markup: LaTeX (including Ti*k*Z, beamer, BibTeX styles); HTML & CSS; Markdown; pandoc.

Other: Unix-like operating systems (macos, Ubuntu Linux, Arch Linux); git & GitHub.