Overview: Doctoral student aiming to rigorously understand catastrophic and existential risk from AI. Experience at Timaeus, Krueger AI Safety Lab, and CHAI. Background in the theory of computation and learning, empirical deep learning, and software engineering. Passionate and prolific educator.

§ Education

Doctor of Philosophy in Computer Science Oct 2024–present
University of Oxford (Department of Computer Science & Magdalen College)
Understanding neural mechanism emergence (e.g., 1). Supervised by Professor Alessandro Abate.

Awards: *Clarendon Scholarship.*

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 neural network geometry (13) leading to three papers (11; 12; 10). Supervised by Daniel Murfet. Founded an AI safety reading group.

Awards: *Dean's Honours List* (top 5% marks in School of Engineering). *Top thesis mark* since degree was first conferred in 2021. *Melbourne School of Engineering Foundation Scholarship.*

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 (unawarded).

§ Research Experience

Research Mentor Oct 2025–present
Research Impact Oxford (RIO) & Independent
As a RIO mentor, I scoped out and mentored a term-long research project for a team of impact-oriented undergraduate students. After the competition, the project has continued independently (paper in preparation). Alongside RIO, I independently formed a team of master's students, scoped out and mentored another research project leading to a workshop paper (2).

Awards: *1st place prize* (best student project, RIO Michaelmas 2025).

Research Affiliate Sep 2025–present
Timaeus (non-profit research organisation)
Advising theory/empirical research generalising singular learning theory to the reinforcement learning setting, leading to a journal paper under review (3).

Research Associate Aug 2023–Sep 2024
Timaeus (non-profit research organisation)
Research engineer and writing lead for papers on the emergence of in-context learning in transformers (7; 6; 5), later leading to a position paper (4).

Research Assistant (AI alignment & reward hacking) Sep 2023–Oct 2024
Krueger AI Safety Lab & Computational and Biological Learning Lab, University of Cambridge
Led project on understanding and mitigating goal misgeneralisation in deep reinforcement learning, including proving game-theoretic performance guarantees and implementing deep reinforcement learning algorithms in JAX, leading to a conference paper (8).

Research Assistant (Human-agent interaction) Jan–Jul 2023
School of Computing and Information Systems, the University of Melbourne
Assisted with explainable AI research project. Automated Qualtrics survey generation with Python.

Research Intern Jun–Oct 2021
Center for Human-compatible AI, University of California (Berkeley)
Theoretical research on limits of reward learning, leading to conference paper (9). Initiated a virtual mini-conference for interns to share presentations about their projects.

§ Teaching Experience (Selected)

Author and Teaching Assistant (Deep Learning Theory) Apr 2026–present
Iliad (non-profit academic organisation)
Creating lecture notes (14). Delivering one-day courses on singular learning theory.

YouTube Lecturer (JAX for Deep Learning Research) Dec 2025–Jan 2026
Independent
Created 13-hour open-source lecture course on introductory JAX for deep learning research (15).

Head Teaching Assistant (AI Safety & Alignment) Oct 2025
Autonomous Intelligent Machines and Systems EPSRC CDT, University of Oxford
Coordinated teaching team for Fazl Barez's *AI Safety & Alignment* graduate course. Developed notebook tutorial on reward hacking and goal misgeneralisation.

Teaching Assistant and Guest Lecturer (AI Ethics) 2021, 2023–2024
Centre for AI and Digital Ethics, the University of Melbourne
Facilitating discussions, marking essays in the ethics and governance of AI for technical graduate students. Guest lecture on ethical perspectives on the race to superintelligence (16).

Lecturer, Head TA, TA (Computer Science) 2016–2021 (full time 2017–2018)
School of Computing and Information Systems, the University of Melbourne
Lecturer for introductory Python course: created lectures and assessments (150 students). Head TA for courses in algorithmics; theoretical computer science, AI: designed problem sets, projects, exams; coordinated fair assessment (300–600 students/course); pioneered digital teaching/assessment methods (17; 18). TA for above courses plus operating systems, networks, and security.

Awards: *School of Engineering Tutor Community Excellence Award (finalist). School of Engineering Most Innovative Academic (finalist). Head Tutor Special Commendation Award. School of Computing and Information Systems Excellence in Tutoring Award.*

Freelance Tutor, Lecturer, Author (Secondary Mathematics & Physics) 2014–2017
Independent & through various schools, tutoring companies, university programs
Private tutoring services. Small-group tutoring services for *Mount Lilydale Mercy College* and *Scotch College, Melbourne.* With *ATAR Notes,* delivered accessible lectures/webinars and published a comprehensive set of course notes. With Mount Lilydale Mercy College, initiated an annual final-year exam revision lecture series, including recruiting high-achieving alumni lecturers.

§ Professional Service and Community Building

Executive Committee Member Mar 2025–present
Oxford AI Safety Initiative
Setting community strategy (with president and other executive committee members). Organising public lectures and other events. Founding and directing a weekly AI safety research seminar.

Reviewer (Machine Learning Conferences) Jun 2024–present
NeurIPS 2024. ICLR 2025. ICML 2025. NeuS 2025, 2026.

§ Technical Proficiencies

Programming
Python (including NumPy, SciPy, matplotlib, PyTorch, PyTorch/XLA, einops, JAX); Haskell (including alex, parsec); C (including sockets); Java; Bash; JavaScript (including React); Rust.

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

Systems
Unix-like operating systems (including macOS, Ubuntu Linux, Arch Linux); administering multi-user VPS; basic Google Cloud Platform (including GCS buckets, TPU VMs).

Other
(Neo)Vim; Make; Git; W&B; Claude Code.

See also: select personal projects on **GitHub** (@matomatical).

§ Publications

*Equal contribution.

Science of deep learning

(1) **MFR,** "Structure and Scale in Simplicial Sequence Modelling." *HiLD: 4th Workshop on High-dimensional Learning Dynamics,* **ICML 2026 workshop.** Preprint in preparation.

(2) Afiq Abdillah Effiezal Aswadi*, Oliver Britton*, Ross Baker*, **MFR,** "Temporal task diversity: Inductive bias under non-stationarity in synthetic sequence modelling." *FoGen: Foundations of Deep Generative Models,* **ICML 2026 workshop.** Preprint arXiv:2605.18281.

(3) Chris Elliott*, Einar Urdshals*, David Quarel*, **MFR,** Daniel Murfet, "Stagewise reinforcement learning and the geometry of the regret landscape." **2026.** Preprint arXiv:2601.07524.

(4) Simon Pepin Lehalleur*, Jesse Hoogland*, **MFR*,** Susan Wei, Alexander Gietelink Oldenziel, George Wang, Liam Carroll, Daniel Murfet, "You are what you eat: AI alignment requires understanding how data shapes structure and generalisation." **2025.** Preprint arXiv:2502.05475.

(5) Liam Carroll, Jesse Hoogland, **MFR,** Daniel Murfet, "Dynamics of transient structure in in-context linear regression transformers." **2025.** Preprint arXiv:2501.17745.

(6) Jesse Hoogland*, George Wang*, **MFR,** Liam Carroll, Susan Wei, Daniel Murfet, "Loss landscape degeneracy and stagewise development in transformers." **TMLR 2024.** Preprint arXiv:2402.02364.

(7) George Wang*, **MFR*,** Jesse Hoogland, Liam Carroll, Susan Wei, Daniel Murfet, "Loss landscape geometry reveals stagewise development of transformers." *HiLD: 2nd Workshop on High-dimensional Learning Dynamics,* **ICML 2024 workshop.** OpenReview (poster).

Awards: *Best papers of HiLD award.*

Reward ambiguity and generalisation in reinforcement learning

(8) Karim Abdel Sadek*, **MFR*,** Usman Anwar, Hannah Erlebach, Christian Schroeder de Witt, David Krueger, Michael Dennis, "Mitigating goal misgeneralization via minimax regret." **RLC 2025.** OpenReview (poster).

(9) Joar Skalse*, **MFR*,** Alessandro Abate, Stuart Russell, Adam Gleave, "Invariance in policy optimisation and partial identifiability in reward learning." **ICML 2023.** PMLR (poster, video).

Neural network geometry

(10) **MFR,** "Proximity to losslessly compressible parameters." **2024.** Preprint arXiv:2306.02834.

(11) **MFR,** "Losslessly compressible neural network parameters." *Machine Learning and Compression,* **NeurIPS 2024 workshop.** OpenReview.

(12) **MFR,** "Functional equivalence and path connectivity of reducible hyperbolic tangent networks." **NeurIPS 2023.** OpenReview (poster, video).

(13) **MFR,** *Structural Degeneracy in Neural Networks.* Master's thesis, School of Computing and Information Systems, the University of Melbourne. **2022.** Available online.

Published teaching materials

(14) Kai Ogden, **MFR,** Zach Furman, "Degeneracy in deep learning: An invitation to singular learning theory." Pilot edition, **2026.** Available online.

(15) **MFR,** "Hi, JAX! An introduction to vanilla JAX for deep learning research." Online edition, **2026.** Course information online. Lecture recordings on YouTube.

(16) **MFR,** "Ethics and the Future of Intelligence." Guest lecture, *The Ethics of Artificial Intelligence,* the University of Melbourne. **2024.** Recording online.

Computer science education

(17) **MFR,** Bryn Jeffries, Harald Søndergaard, "Teaching simple constructive proofs with Haskell programs." **TFPIE 2022,** doi:10.4204/EPTCS.363.4.

(18) **MFR,** Bryn Jeffries, Harald Søndergaard, "Programming to learn: Logic and computation from a programming perspective." **ACM ITiCSE 2022,** doi:10.1145/3502718.3524814.