Logan Graham

[redacted] - LinkedIn - logangraham.xyz - GitHub

ACADEMIC

D.Phil (PhD) in Engineering Science, University of Oxford 2015-2020

- Recipient of the 2015 Rhodes Scholarship (British Columbia & Balliol)
- Member of the Machine Learning Research Group in the Department of Engineering Science. Produced a PhD on machine learning, causality, and interpretability.

Bachelor of Arts in Economics (Hons.), University of British Columbia 2011-2015

- Awarded a National Scholar, Wesbrook Scholar, and C.K. Choi Scholar – the university's three most prestigious undergraduate awards

EXPERIENCE

Anthropic San Francisco

Head of the Frontier Red Team Nov 2022 - present

- Senior lead responsible for evaluating and enhancing advanced capabilities of AI models
- Designed, directed, and built the first system to test whether the most advanced AI models may present national security risks
- Led red teaming, evaluation, and model capabilities research projects in cybersecurity, biosecurity, and advanced model autonomy
- Built and directed a team of 15 researchers and dozens of external contractors
- Brief and communicate our work frequently to allied governments and top global media (features in *60 Minutes, the Wall Street Journal, Washington Post*, and more)

10 Downing Street, the Prime Minister's Office London, UK

Special Adviser to the Prime Minister July 2020 - July 2022

- Lead adviser to the Prime Minister on artificial intelligence, £55bn of R&D funding, technology security, and data policy.
- Regularly briefed the Prime Minister, Chief of Staff, Ministers, Generals, and Secretaries of State on technology, and orchestrated meetings between FAANG and unicorn CEOs
- Particular achievements in AI security and policy: conceived of and wrote the UK National AI Strategy, started the $1bn compute review, established the UK's AI talent recruitment program, and won $20m in extra funding for AI masters and PhDs.
- Co-led science and technology policy. Oversaw £55bn of R&D funding, established a $1bn DARPA-style science agency, founding data scientist of the Prime Minister's Data Science team, successfully argued for a £5bn uplift in R&D funding; oversaw the Defence AI Strategy, International Technology Strategy, Biosecurity Strategy, and Digital Strategy.

Babylon Health London, UK

Research Scientist July 2019 - Oct 2019

- Published 2 papers at a top journal and top conference workshop as lead researcher on project investigating causality and counterfactual inference for safe, interpretable, efficient machine learning
- Built and open-sourced *Twin Networks*, a 2,000 line-of-code Python library for handling Structural Causal Models and performing probablistic inference on them
- Second engineer and paper lead on *Multiverse*, an open-source probabilistic programming engine for approximate inference with structural causal models

X (formerly Google[X]) Mountain View, California & London, UK

AI Resident & Consultant Apr 2019 - Jan 2020

- First hire and computational modeling lead for a machine learning + biology "moonshot" project inside X, Google's research lab. Project received several million $ in follow-on funding and now has 10 team members as a major X project
- Wrote the first 5,000 lines of code in Python, TensorFlow, and internal Google libraries. Built an end-to-end experiment pipeline to analyse high-dimensional genomics data at high performance on Google infrastructure

- Created a deep Bayesian model that matched the state of the art in computational agriculture. Resulted in 1 patent and a Department of Energy grant
- Hired and led a team of three machine learning researchers who ultimately developed models that exceeded state of the art

Children's Arthritis Foundation Vancouver, Canada
Co-founder 2001-2015

- Raised over $150,000 speaking directly to wealthy donors and organising fundraisers. Provoked change in medical legislation through speaking to the governing political party at 12 years old.
- Established foundation at 7 with my family after being diagnosed with Rheumatoid Arthritis at 4.

TECHNICAL CAPABILITIES

- Scientific computing (Python, R); Probabilistic Machine Learning (TensorFlow, PyTorch, Keras, Pyro, PyMC, scikit-Learn, SciPy stack); data science (plotly, dash, pandas, matplotlib)

COMMUNITY & TECHNOLOGY

- Fellow, Interact (2017-present); admissions panel member (2022). *Interact (joininteract.com) is a VC-funded community of young technologists committed to using technology to improve the world. It has a <10% acceptance rate and includes several unicorn founders, including companies like Scale AI and Opendoor.*
- Fellow, South Park Commons (2021-2023). *South Park Commons (southparkcommons.com) is a community of technologists dedicated to building new technology. It has a <10% acceptance rate.*

PUBLICATIONS

M. Sharma, ..., **L. Graham**, et al. "Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming" *ArXiv preprint*, 2024.

E. Hubinger, ..., **L. Graham**, et al. "Sleeper agents: Training deceptive llms that persist through safety training" *ArXiv preprint*, 2024.

Y. Perov*, **L. Graham***, K. Gourgoulias, J. G. Richens, C. M. Lee, A. Baker, S. Johri, "MultiVerse: Causal Reasoning using Importance Sampling in Probabilistic Programming," *Symposium on Advances in Approximate Bayesian Inference*. Proceedings of Machine Learning Research, 2020. (Invited poster presentation, and 2020 talk at the University of Toronto.)

M. Brundage, ..., **L. Graham**, et al. "Toward trustworthy AI development: mechanisms for supporting verifiable claims." *arXiv preprint*, 2020.

M. Willis, P. Duckworth, **L. Graham**, M. Osborne, E. Meyer, A. Coulter. "The Future of Healthcare: Computerisation, Automation, and General Practice Services," *The Health Foundation*, 2020.

L. Graham, A. Gilbert, J. Simons, A. Thomas. "Artificial intelligence in hiring: Assessing impacts on equality." *The Institute for the Future of Work*, 2020.

L. Graham, C. M. Lee, Y. Perov, "Copy, paste, infer: A robust analysis of twin networks for counterfactual inference," *33rd Neural Information Processing Systems Workshop in Causal Machine Learning*, 2019.

P. Duckworth*, **L. Graham***, M. Osborne, "Inferring Work Task Automatability from AI Expert Evidence," *2nd AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society*, 2019. Invited talk at

NeurIPS 2018 AI for Social Good workshop, link: https://www.youtube.com/watch?v=DXjE1YVKA40

PhD THESIS **L. Graham**, "Interpretable causal systems: interpretability and causality in machine learning for human and nonhuman decision-making ," *Doctor of Philosophy Thesis*, 2020.
https://ora.ox.ac.uk/objects/uuid:9b95b73d-34a9-42c8-b2d7-11efbb15609a