Ryan Carey, PhD

Education

University of Oxford *Ph. D. in Statistics* Oct 2020 - Oct 2024
- *Incentives and AI Safety in Causal Models*, accepted without corrections

Imperial College London *MSc. Theoretical Systems Biology* Oct 2014 - 2015

Monash University *Bachelor of Medicine / Surgery w. Distinction* 2008 - 2012

Employment

Millennium Management *Quantitative Researcher* Jul 2024 - Dec 2024
Quantitative Research Intern Jan 2024 - Apr 2024

DeepMind *Research Intern* Jun 2022 - Oct 2022

University of Oxford *Research Fellow* Nov 2018 - Jun 2022
- Team lead for technical AI safety work at the Future of Humanity Institute

OpenAI *Research Engineering Intern* Mar 2018 - Apr 2018
- Worked with Paul Christiano on evaluating SAT problems using transformer models

Contracting with Ought, BERI *Research Engineer* rest of Oct 2017 - Oct 2018

Machine Intelligence Research Institute *Assistant Research Fellow* Sep 2016 - Oct 2017

Monash Health, and other hospitals *Medical Resident* Aug - Sep 2014, Jul 2015- May 2016
- Contract/locum work, <50% of full-time, mostly in emergency medicine

Monash Health *Medical Intern* Jan 2013 - Dec 2013

Academic Publications

- *Invariances in the Selected Margins of Causal Bayesian Networks* **R. Carey**, M. M. Ansanelli, E. Wolfe, R. Evans.
- *Incentives for Responsiveness, Instrumental Control and Impact* **R. Carey**, E. Langlois, C. van Merwijk, T. Everitt. (in review @AIJ)
- *Toward a Complete Criterion for Value of Information*. **R. Carey**, S. Lee, E. Bareinboim, R. Evans. TMLR, 2024.
- *Human Control: Definitions and Algorithms*. **R. Carey**, T. Everitt. UAI, 2023.
- *Reasoning about Causality in Games*. L. Hammond, J. Fox, T. Everitt, **R. Carey**, A. Abate, M. Wooldridge. AIJ, 2023.
- *Path-specific Objectives for Safer Agent Incentives*. S. Farquhar, **R. Carey**, T. Everitt. AAAI, 2022.
- *A Complete Criterion for Value of Information in Soluble Influence Diagrams*. C. van Merwijk*, **R. Carey***, T. Everitt. AAAI, 2022.
- *Why Fair Labels Can Yield Unfair Predictions: Graphical Conditions for Introduced Unfairness*. C. Ashurst, **R. Carey**, S. Chiappa, T. Everitt. AAAI, 2022.
- *PyCID: A Python Library for Causal Influence Diagrams*. J. Fox, T. Everitt, **R. Carey**, E. Langlois, A. Abate, M. Wooldridge, Scipy, 2021.
- *Agent Incentives: a Causal Perspective*. T. Everitt*, **R. Carey***, E. Langlois*, P. A. Ortega, S. Legg. AAAI, 2021.
- *(When) Is Truth-telling Favored in AI Debate?*. V. Kovarik, and **R. Carey**. AAAI workshop, 2020.

- *Incorrigibility in the CIRL Framework*. **R. Carey**. AIES, 2018.
- *Predicting Human Deliberative Judgments with Machine Learning*. O. Evans, A. Stuhlmuller, C. Cundy, **R. Carey**, Z. Kenton, T. McGrath, A. Schreiber, Technical report, University of Oxford, 2018.

AI Safety-related Projects and Service

2020-2024 **Cofounder of the Causal Incentives Working Group**.

- A collection of researchers from DeepMind, Oxford, UChicago, Imperial, etc. focused on using causal models to understand agency and AI risk.

2014 **Founded Effective Altruism Forum**

- I cofounded, and operated the main website for online effective altruism discussion for several years. It has since been adopted by the Centre for Effective Altruism, has 10 million views annually, and is now managed by several full-time staff.

Teaching & Invited Presentations

2023 Towards Causal Foundations of Safe AI, **invited talk, Interactive Causal Learning Conference**
2023 Towards Causal Foundations of Safe AI, **tutorial co-presenter, UAI**
2023 Statistical Programming, **teaching assistant, Oxford**
2022 Probability Theory, **teaching assistant, Oxford**
2019 **DeepMind** Safety Seminar, *The Incentives that Shape Behaviour*
2019 **DeepMind** Iceland AGI Safety Workshop, *The Incentives that Shape Behaviour*
2018 Center for Human-compatible AI, **UC Berkeley**, *Incorrigibility in the CIRL Framework*