Ryan Carey

Education

University of Oxford *Ph. D. in Statistics* 2020 - (now ready to graduate)
- Cofounded the Causal Incentives Working Group, which has members from DeepMind and 5 top universities. 2021-present

Imperial College London *MSc. Theoretical Systems Biology* 2014 - 2015
Monash University *Bachelor of Medicine / Surgery w. Distinction*

Employment

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

Collabs. with University of Oxford, Ought, BERI *Engineering, Hiring* (most of) Oct 2017 - Oct 2018
OpenAI *Research Engineering Intern* Mar 2018 - Apr 2018
Machine Intelligence Research Institute *Assistant Research Fellow* 2016 - 2017
Monash Health, and other hospitals *Medical Intern/Resident* 2013 - 2014, 2015-2016

AI Safety– and EA–related Projects and Service

2020 **Cofounded 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 **Curated The Effective Altruism Handbook**.
- A collection of EA essays that preceded "Doing Good Better"; thousands of copies circulated, and it is cited on the last page of Derek Parfit's "On What Matters")

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.

2008 **Founded the Felicifia Forum**
- An online utilitarianism forum that followed a blog of the same name; a predecessor of the EA community.

2021-2023 **AI Safety subject matter expert for 80,000 Hours**
2021-2023 **Advisor to Concordia Research**

Academic Publications

- *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?*. Kovarik, Vojtech, and **Carey, Ryan**. AAAI workshop, 2020.
- *Incorrigibility in the CIRL Framework*. **Carey, Ryan**. AIES, 2018.
- *Predicting Human Deliberative Judgments with Machine Learning*. Evans, O. , Stuhlmuller, A. , Cundy, C. , **Carey, R.** , Kenton, Z. , McGrath, T. , & Schreiber, A. Technical report, University of Oxford, 2018.

Manuscripts in preparation

- *Toward a Complete Criterion for Value of Information*. **R. Carey**, S. Lee, E. Bareinboim, R. Evans
- *Incentives in Causal Models* **R. Carey**, E. Langlois, C. van Merwijk, T. Everitt.
- *Selection and Marginalisation in DAGs* **R. Carey**, M. M. Ansanelli, E. Wolfe, R. Evans.

Research Blog Posts on Altruistic Strategy (prepared for various EA organisations)

- *Interpreting AI Compute Trends*. **R. Carey**. AI Impacts. 2018.
- *The Payoff and Probability of Obtaining Venture Capital*. **R. Carey**. 80,000 Hours Blog. 2014
- *What is the Average Net Worth of Stanford Entrepreneurs?*. **R. Carey**. 80,000 Hours Blog. 2014
- *How much do YCombinator Founders Earn?*. 80,000 Hours Blog. **R. Carey**. 2014

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*
2019 **Guest Lecture** for Oxford's Centre for Doctoral Training, *Incentives and AI Safety*
2018 Center for Human-compatible AI, **UC Berkeley**, *Incorrigibility in the CIRL Framework*

Accomplishments

2023 **Awarded $212,000 from Long-term Future Fund and Lightspeed Grants for AI Safety research**
2023 **Top 10% Reviewer at AIStats**
2019 **Alignment Prize (for work on corrigibility), $2,500**