Ryan Carey

causalincentives.com • RyanCarey

Education

University of Oxford *Doctor of Philosophy in Statistics (with A/Prof Robin Evans)* **2020 - present**
- Cofounded the Causal Incentives Working Group 2021-present

Imperial College, London *Master of Science in Bioinformatics and Theoretical Systems Biology* **2014 - 2015**

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

Employment

DeepMind *Research Intern* **Jul 2022 - Oct 2022**

University of Oxford, Future of Humanity Institute *Research Fellow* **Nov 2018 - Jul 2022**
- Team lead for AI safety 2019-2022

OpenAI *Research Engineering Intern* **2018**

University of Oxford, Future of Humanity Institute *Research Intern* **2016 - 2017**

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

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*, PA. Ortega, S. Legg. AAAI, 2021.
- *(When) Is Truth-telling Favored in AI Debate?*. Kovarik, Vojtech, and **Carey, Ryan**. SafeAI@AAAI, 2020.
- *How useful is quantilization for mitigating specification-gaming?*. **Carey, Ryan**. ICLR workshop, 2019.
- *Incorrigibility in the CIRL Framework*. **Carey, Ryan**. Proceedings of 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.

Teaching & Invited Presentations

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 **Oxford's Centre for Doctoral Training**, *Incentives and AI Safety* [lecture and tutorial]
2018 **Center for Human-compatible AI, UC Berkeley**, *Incorrigibility in the CIRL Framework*
2018 **AAAI/ACM Conference on AI, Ethics, and Society**, *Incorrigibility in the CIRL Framework*

Awards

2023 **Top 10% Reviewer at AIStats**
2019 **Alignment Prize: Second Prize, $2,500**