# Ryan Carey, PhD

## Education

**University of Oxford** *Ph. D. in Statistics* 2020 - Oct 2024

- *Incentives and AI Safety in Causal Models*, accepted without corrections

**CFA UK** *Investment Management Certificate* Oct 2024

**Imperial College London** *MSc. Theoretical Systems Biology*

**Monash University** *Bachelor of Medicine / Surgery w. Distinction*

## Employment

**Millennium Management** *Quantitative Researcher* Jul 2024 - Dec 2024

*Quantitative Research Intern* Jan 2024 - Apr 2024

- I researched reversion, intraday, and vendor-data alphas in a fully systematic mid-frequency equities pod.
- This revealed a signal with Sharpe $2.0$, that improved model performance by $\$1.5M/yr$ on backtest.
- I productionised this signal by writing signal generation and model-fitting code in Python.
- "Ryan joined our team as an intern in January 2024, researching some signals for our stat arb book, and performed well, leading us to make him a full time offer when he finished in April. He joined full time in the summer after finishing his PhD and did some further good work for the same strategies. Unfortunately the pod was shut down in October, certainly through no fault of Ryan's. I strongly recommend him as a potential quantitative researcher or trader candidate." —— Charles Dillon (Senior Portfolio Manager)

**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 deep learning

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

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

## 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.

## 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*

## Other

- Offered $212,000 grant from Long-term Future Fund and Lightspeed Grants for AI research 2023
- Awarded $2,500 Alignment Prize for work on AI safety 2019
- Quantitative Trader job offer from Jane Street 2016
- $> \$3k$ net profit in online Texas Hold'em, without depositing real money, and $> \$1k$ in sports betting arbitrage