Ryan Carey

careyryan.com • RyanCarey

Research Positions

University of Oxford, Future of Humanity Institute **Nov 2018 - Present**

Research Fellow

- Sole-authored and published *Incorrigibility in the CIRL Framework*.
- Team lead for the AI safety team from *Q4 2019* onwards.
- Supervised a visiting postdoc's project: *(When) is Truth-telling Favored in AI Debate*.
- *External collaborator* at **Google DeepMind**, from **July 2019 to Present**.
  - Co-lead-authored *The Incentives that Shape Behavior*.
  - Two manuscripts in-progress with DeepMind research scientist Tom Everitt: *Understanding Agent Incentives using Causal Influence Diagrams: Multi-decision scenarios with sufficient recall*, and (lead author of) *Mediator Problems in AI safety*.

University of Oxford, Future of Humanity Institute **Nov 2017 - Jul 2018**

Research Intern

- Lead engineer for *Predicting Human Deliberative Judgments* project (in collaboration with Ought. Inc).

OpenAI **Feb 2018 - Mar 2018**

Research Engineering Intern

- Designed neural networks to heuristically decompose SAT problems, with researcher Paul Christiano.

Machine Intelligence Research Institute **Jul 2016 - Sep 2017**

Assistant Research Fellow

- Research blog post *Addressing Three Problems with Counterfactual Corrigibility* won a $\$2,500$ "Alignment Prize".

Publications

- *The Incentives that Shape Behavior*
  Carey, R, Langlois, L, Everitt, T & Legg, S. SafeAI@AAAI, 2020.
- *(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. Workshops at International Conference on Learning Representations, 2019.
- *Incorrigibility in the CIRL Framework*
  Carey, Ryan. Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. ACM, 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.

Invited Lectures and Presentations

2019 **DeepMind Safety Seminar**, *The Incentives that Shape Behavior*
2019 **DeepMind Iceland AGI Safety Workshop**, *The Incentives that Shape Behavior*
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*

Education

Imperial College, London **2014 - 2015**

Masters of Science in Bioinformatics and Theoretical Systems Biology

Monash University **2008 - 2012**

Bachelor of Medicine / Bachelor of Surgery Distinction