Owain Evans

Curriculum Vitae

Research Scientist in Machine Learning, working on how to make AI safe and beneficial.

Education

2008–2015 **PhD in Philosophy**, *Massachusetts Institute of Technology*.
Supervisors: Roger White (philosophy of science), Vikash Mansinghka (machine learning).

2004–2008 **BA in Philosophy and Mathematics**, *Columbia University*.

Employment

2017–now **Research Scientist**, *Future of Humanity Institute, University of Oxford*.
Machine Learning research focused on AI Safety: learning human preferences, safe RL, and active learning.

2015–2017 **Postdoctoral Researcher**, *Future of Humanity Institute, University of Oxford*.

2013–2015 **Research Assistant**, *MIT Probabilistic Computing Project, Massachusetts Institute of Technology*.

Publications

[1] Zachary Kenton, Angelos Filos, Owain Evans, and Yarin Gal. Generalizing from a few environments in safety-critical reinforcement learning. In *Safe ML, ICLR Workshop*, 2019.

[2] Owain Evans, William Saunders, and Andreas Stuhlmüller. Machine learning projects for iterated distillation and amplification. Technical report, 2019.

[3] Owain Evans, Andreas Stuhlmüller, Chris Cundy, Ryan Carey, Zachary Kenton, Thomas McGrath, and Andrew Schreiber. Predicting human deliberative judgments with machine learning. Technical report, 2018.

[4] Sebastian Schulze and Owain Evans. Active reinforcement learning with monte-carlo tree search. *arXiv preprint arXiv:1803.04926*, 2018.

[5] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The

16/17 St Ebbe's Street – Oxford OX1 1PT – United Kingdom
+447525200961 • *owain.evans@philosophy.ox.ac.uk*
owainevans.github.io • *owain-evans-78b210133* • *owainevans*

malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

[6] William Saunders, Girish Sastry, Andreas Stuhlmueller, and Owain Evans. Trial without error: Towards safe reinforcement learning via human intervention. *arXiv preprint arXiv:1707.05173*, 2017.

[7] Katja Grace, John Salvatier, Allan Dafoe, Baobao Zhang, and Owain Evans. When will AI exceed human performance? Evidence from AI experts. *arXiv preprint arXiv:1705.08807*, 2017.

[8] David Krueger, Jan Leike, Owain Evans, and John Salvatier. Active reinforcement learning: Observing rewards at a cost. In *Future of Interactive Learning Machines, NIPS Workshop*, 2016.

[9] Owain Evans, Andreas Stuhlmüller, and Noah D Goodman. Learning the preferences of ignorant, inconsistent agents. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 323–329. AAAI Press, 2016.

[10] Owain Evans and Noah D Goodman. Learning the preferences of bounded agents. In *NIPS Workshop on Bounded Optimality*, volume 6, 2015.

[11] Owain Evans, Leon Bergen, and Joshua Tenenbaum. Learning structured preferences. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32, 2010.

[12] Tomer Ullman, Chris Baker, Owen Macindoe, Owain Evans, Noah Goodman, and Joshua B Tenenbaum. Help or hinder: Bayesian models of social goal inference. In *Advances in neural information processing systems*, pages 1874–1882, 2009.

Presentations

2018 **Oxford University Psychology Society**, *DeepDream and Seeing As*.

2018 **Creative AI London**, *DeepDream and Seeing As*.

2017 **NIPS 2018, Long Beach CA**, *Predicting Slow Judgments*.

2017 **EA Global London**, *Careers in AI Safety*.

2017 **ETH Zürich Workshop on AI Safety**, *Trial Without Error*.

2017 **Center for Future of Intelligence, Cambridge**, *Trial Without Error*.

2017 **University College London Machine Learning**, *Trial Without Error*.

2017 **Deepmind-FHI AI Safety Seminar**, *Trial Without Error*.

2017 **Oxford University Machine Learning Workshop**, *Trial Without Error*.

2017 **Asilomar Conference on Beneficial AI**, *Learning the Preferences of Ignorant, Inconsistent Agents*.

2017 **AAAI 2017, Phoenix AZ (oral)**, *Learning the Preferences of Ignorant, Inconsistent Agents*.

2017 **AAAI 2017, Phoenix AZ (Ethics Workshop)**, *agentmodels.org*.

2016 **University of Toronto Machine Learning**, *Trial Without Error*.

16/17 St Ebbe's Street – Oxford OX1 1PT – United Kingdom
+447525200961 • *owain.evans@philosophy.ox.ac.uk*
owainevans.github.io • *owain-evans-78b210133* • *owainevans*

2016 **Atomico European AI Vanguard**, *Learning the Preferences of Ignorant, Inconsistent Agents*.

2016 **Oxford TORCH Humanities Centre**, *Automated Corporations and AI Risk*.

2016 **EA Global Oxford**, *Careers in AI Safety*.

2016 **Effective Altruism Berkeley**, *Learning Human Preferences*.

2015 **Oxford University Probabilistic Programming Group**, *Learning Human Preferences*.

2015 **Stanford University Computational Cognitive Science**, *Learning Human Preferences*.

2014 **DARPA Summer School on Probabilistic Programming**, *Intro to Probabilistic Programming in Venture*.

2014 **Cambridge University Machine Learning Group**, *Intro to Probabilistic Programming in Venture*.

2014 **Oxford University Machine Learning**, *Intro to Probabilistic Programming in Venture*.

2010 **Cognitive Science Society Conference 2010**, *Learning Structured Preferences*.

Grants

2015-2018 **Future of Life Institute**, *Inferring Human Values*, $227K.

Teaching

2014 **DARPA Summer School on Probabilistic Programming**, *Portland OR*.

2014 **Tutorial on Probabilistic Programming**, *Cambridge, UK*.

2013 **Paradox and Infinity Undergraduate Course**, *MIT, USA*.

2010 **Intro to Political Philosophy**, *MIT, USA*.

16/17 St Ebbe's Street – Oxford OX1 1PT – United Kingdom
+447525200961 • *owain.evans@philosophy.ox.ac.uk*
owainevans.github.io • *owain-evans-78b210133* • *owainevans*