# Owain Evans

*Curriculum Vitae*

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

## Employment

2020–now **Research Associate**, *Future of Humanity Institute, University of Oxford*.
AI Safety, Truthful/Honest AI

2019–2020 **Research Scientist**, *Ought, San Francisco*.
AI Safety research on amplification (both ML and human experiments) and tools for forecasting (e.g. AI timelines)

2017–2019 **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] Andy Zou, Tristan Xiao, Ryan Jia, Joe Kwon, Mantas Mazeika, Richard Li, Dawn Song, Jacob Steinhardt, Owain Evans, and Dan Hendrycks. Forecasting future world events with neural networks. *arXiv preprint arXiv:2206.15474*, 2022.

[2] Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models, 2022.

[3] Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*, 2022.

[4] Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. Truthful AI: Developing and governing AI that does not lie. *arXiv preprint arXiv:2110.06674*, 2021.

[5] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, 2022.



*18 Talbot Street – Cardiff CF119BW – United Kingdom*
*+14152319736* • *owaine@gmail.com* • *owainevans.github.io*
*owain-evans-78b210133* • *OwainEvans_UK*

[6] Mihaela Curmei, Andrew Ilyas, Owain Evans, and Jacob Steinhardt. Constructing and adjusting estimates for household transmission of sars-cov-2 from prior studies, widespread-testing and contact-tracing data. *International journal of epidemiology*, 50(5):1444–1457, 2021.

[7] Tim Colbourn, William Waites, Jasmina Panovska-Griffiths, David Manheim, Simone Sturniolo, Greg Colbourn, Cam Bowie, Keith M Godfrey, Julian Peto, Rochelle A Burgess, et al. Modelling the health and economic impacts of population-wide testing, contact tracing and isolation (ptti) strategies for covid-19 in the uk. 2020.

[8] Mihaela Curmei, Andrew Ilyas, Owain Evans, and Jacob Steinhardt. Estimating household transmission of sars-cov-2. *medRxiv*, 2020.

[9] Saunders, William and Rachbach, Ben and Evans, Owain and Miller, Zachary and Byun, Jungwon and Stuhlmüller, Andreas. Evaluating arguments one step at a time. `https://ought.org/updates/2020-01-11-arguments`, 2020. Accessed 11-January-2020.

[10] Owain Evans. Sensory optimization: Neural networks as a model for understanding and creating art. *arXiv preprint arXiv:1911.07068*, 2019.

[11] 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.

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

[13] 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.

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

[15] Miles Brundage, Shahar Avin, Jack Clark, Helen Toner, Peter Eckersley, Ben Garfinkel, Allan Dafoe, Paul Scharre, Thomas Zeitzoff, Bobby Filar, et al. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228*, 2018.

[16] 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.

[17] 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.

[18] 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.

*18 Talbot Street – Cardiff CF119BW – United Kingdom*
*+14152319736* • *owaine@gmail.com* • *owainevans.github.io*
*owain-evans-78b210133* • *OwainEvans_UK*

[19] 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.

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

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

[22] 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.

## 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*.

## Presentations

2022 **Anthropic, San Francisco**, *Teaching models to express uncertainty in words*.

2022 **CHAI Workshop on AI Safety**, *Teaching models to express uncertainty in words*.

2022 **ACL 2022 Conference in Dublin**, *TruthfulQA*.

2022 **OATML University of Oxford**, *Teaching models to express uncertainty in words*.

2022 **DeepMind-FHI AI Safety Seminar**, *TruthfulQA*.

2021 **OATML University of Oxford**, *TruthfulQA*.

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*.

*18 Talbot Street – Cardiff CF119BW – United Kingdom*
*+14152319736* • *owaine@gmail.com* • *owainevans.github.io*
*owain-evans-78b210133* • *OwainEvans_UK*

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

2018-2021 **Future of Life Institute**, *Factored Cognition: Amplifying Human Cognition for Safely Scalable AGI (w/ Andreas Stuhlmueller)*, $225K.

2015-2018 **Future of Life Institute**, *Inferring Human Values (w/ Andreas Stuhlmueller)*, $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*.

*18 Talbot Street – Cardiff CF119BW – United Kingdom*
*+14152319736* • *owaine@gmail.com* • *owainevans.github.io*
*owain-evans-78b210133* • *OwainEvans_UK*