☎ *016502509794*
✉ *chris.j.cundy@gmail.com*
C-J-Cundy
Updated May 17, 2025

Chris Cundy

Experience

June 2024– **Research Scientist**, *FAR.AI*, Berkeley, California, USA

Characterising and mitigating catastrophic risks from frontier AI systems. My main responsibilities are:

- Leading research projects and determining the direction for teams of researchers and engineers.
- Designing and implementing training and evaluation schemes with state-of-the-art models, conducting empirical studies, and communicating results through papers and presentations.
- As a research lead, ensuring that research aims are consistent FAR's mission.

June 2022–September 2022 **Research Scientist Intern**, *Technical AI Safety Team*, DeepMind, London, UK

Investigating robust and reliable machine learning in theory and at scale

- Investigated susceptibility of autoregressive models to delusions, where unobserved latent variables lead to incorrect probabilistic judgments.
- Developed a theoretical model for delusions; investigated delusions at scale by analysing performance of DeepMind's Gato (a large generalist, multi-task autoregressive model) on custom environments.

October 2017–January 2018 **Visiting Scholar**, *Future of Humanity Institute*, University of Oxford, Oxford, UK

Developing algorithms to predict human judgments. Supervised by Owain Evans and Andreas Stuhlmüller

- Designed algorithms to collate quick, noisy human judgments to predict the answer to complicated tasks which would typically require deliberation.

June–September 2017 **Visiting Scholar**, *Centre for Human-Compatible AI*, University of California, Berkeley, US

Supervised by Daniel Filan & Stuart Russell, researching topics in AI safety

- Extended previous work on inverse reinforcement learning to the options framework for hierarchical reinforcement learners. Formalized the problem, derived theoretical results, performed experiments on data; later presented at an ICML workshop.

Education

2018-2024 **PhD - Computer Science**, *Stanford University*, Stanford, California, USA

- Advised by Stefano Ermon.
- Investigating topics in inverse reinforcement learning, sequence modelling and variational inference.
- Thesis: *Beyond Maximum Likelihood: Distribution-Aware Machine Learning*.

2016-2017 **MEng - Computer Science**, *University of Cambridge*, Cambridge, UK

- Grade: Distinction.
- Modules Include: Data Science in R, Probabilistic Machine Learning, Category Theory, Machine Learning and Algorithms for Data Mining, Social and Technological Network Data Analytics.
- Thesis: *Investigating Variational Gaussian Process State-Space Models with Gaussian Likelihood*. Supervised by Carl E. Rasmussen.

2013-2016 **BA - Natural Sciences (Physics)**, *University of Cambridge*, Cambridge, UK

- Grade: 1st. Modules: Physics, Maths, Chemistry, Computer Science.

Selected Publications

2024 **SequenceMatch: Imitation Learning for Autoregressive Sequence Modelling with Backtracking**,
Chris Cundy, *Stefano Ermon*, ICLR 2024

2024 **Privacy-Constrained Policies via Mutual Information Regularized Policy Gradients**,
Chris Cundy, *Rishi Desai, Stefano Ermon*, AISTATS 2024

2022 **LMPriors: Pre-Trained Language Models as Task-Specific Priors**,
_Kristy Choi_*, **Chris Cundy***, _Sanjari Srivasta, Stefano Ermon_, First Workshop on Foundation Models for Decision Making, NeurIPS 2022

2021 **BCD Nets: Scalable Variational Approaches for Bayesian Causal Discovery**,
Chris Cundy, _Aditya Grover, Stefano Ermon_, NeurIPS 2021

2020 **Flexible Approximate Inference via Stratified Normalizing Flows**,
Chris Cundy, _Stefano Ermon_, UAI 2020

2018 **Parallelizing Linear Recurrent Neural Nets over Sequence Length**,
Eric Martin, **Chris Cundy**, ICLR 2018

Additional Publications

2021 **IQ-Learn: Inverse soft-Q Learning for Imitation**,
Divyansh Garg, Shuvam Chakraborty, **Chris Cundy**, _Jiaming Song, Stefano Ermon_, NeurIPS 2021

2018 **Exploring Hierarchy-Aware Inverse Reinforcement Learning**,
Chris Cundy, _Daniel Filan_, First Workshop on Goal Specifications for Reinforcement Learning, ICML 2018

2017 **Predicting Slow Judgment**,
Owain Evans, Andreas Stuhlmüller, Ryan Carey, Neal Jean, Andrew Schreiber, Girish Sastry, **Chris Cundy**, First Aligned Artificial Intelligence Workshop, NeurIPS 2017

2015 **Simulation Of Plants In Buildings; Incorporating Plant-Air Interactions In Building Energy Simulation**,
Rebecca Ward, Ruchi Choudhary, Christopher Cundy, George Johnson, Allan McRobie, 14th Conference of International Building Performance Simulation Association

Service

2023 **Teaching Assistant–CS228 (Probabilistic Graphical Models)**, _Stanford University_

2022 **Head Teaching Assistant–CS228 (Probabilistic Graphical Models)**, _Stanford University_
Received award for excellence (awarded to top 5% of Teaching Assistants).

2023 **Project Supervisor**, _Supervised Project for Alignment Research (SPAR)_, Stanford AI Alignment
Supervised five undergraduates on a project finding scaling laws in prompt injections.
Presented work at the 7th Center for Human-Compatible AI workshop.

2021 **Project Supervisor**, _Undergraduate Research Program_, Stanford Existential Risk Initiative
Served as supervisor for an undergraduate project on forecasting AI progress.

2020– **Reviewer**
Reviewed for the following conferences: UAI (2020-2022,2024), ICML (2020,2022,2023,2025), ICLR (2021-2025), NeurIPS (2021-2025), AAAI-(Safe and Robust AI track) (2023).

Relevant Awards

March 2024 **Winner, OpenAI Preparedness Challenge**

- One of the top ten submissions for the OpenAI Preparedness Challenge, for submitting _the most unique, while still being probable, potentially catastrophic misuse of the [OpenAI API]_.
- Developed proof-of-concept showing how GPT4-V, and speech-to-text with GPT4, could be used to parse vast amounts of unlabelled surveillance data and find actionable insights that could be used for blackmail or insider trading.
- Prize: $25,000 in OpenAI credits.