Chris Cundy

016502509794
chris.j.cundy@gmail.com
C-J-Cundy
Updated November 1, 2023

Education

2018-2024 **PhD - Computer Science**, *Stanford University*, Stanford, California, USA.
- Advised by Stefano Ermon
- Investigating topics in inverse reinforcement learning, variational inference and large language models
- Graduating early 2024

2016-2017 **MEng - Computer Science**, *University of Cambridge*, Cambridge, UK.
- Grade: Distinction
- Modules Include: Probabilistic Machine Learning, Machine Learning and Algorithms for Data Mining
- Supervised by Carl E. Rasmussen

2013-2016 **BA - Natural Sciences (Physics)**, *University of Cambridge*, Cambridge, UK.
- Grade: 1st
- Modules: Physics (1st, 2nd, 3rd year) Maths (1st, 2nd year); Chemistry, Computer Science (1st Year)

Selected Publications

2023 **SequenceMatch: Imitation Learning for Autoregressive Sequence Modelling with Backtracking**,
Chris Cundy, *Stefano Ermon*, Under Review.

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

2023 **Neural Networks and the Chomsky Hierarchy**,
*Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, **Chris Cundy**, Marcus Hutter, Shane Legg, Joel Veness, Pedro A. Ortega*, ICLR 2023.

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

2021 **Privacy-Constrained Policies via Mutual Information Regularized Policy Gradients**,
Chris Cundy, *Stefano Ermon*, Preprint.

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.

Relevant Research Experience

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 and derived bounds on probabilistic error as a function of distribution shift
- Investigated relevance of delusions at scale by analysing performance of DeepMind's Gato (a large generalist, multi-task autoregressive model) on custom environments designed to induce delusions.

October 2017–January 2018 **Visiting Scholar**, *Future of Humanity Institute*, University of Oxford, Oxford, UK.
Developing algorithms to predict deliberative human judgements
- In collaboration with Owain Evans and Andreas Stuhlmüller at Ought inc.
- Developed the 'Predicting Slow Judgements' dataset, consisting of responses to questions that require in-depth human thought to answer, e.g. legal verdicts.
- Designed algorithms to collate quick, noisy human judgments to solve these deliberative problems with well-calibrated predictions.

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.

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.
Served as supervisor for a group of five undergraduates on an ongoing, technically challenging project studying scaling laws in prompt injections.
Met weekly, setting goals and overall research direction, and ensuring targets were met.
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), ICML (2020,2022,2023), ICLR (2021-2024), NeurIPS (2021-2023), AAAI-(Safe and Robust AI track) (2023-2024)