# CASSIDY LAIDLAW

cassidy_laidlaw@berkeley.edu
cassidylaidlaw.com

## EDUCATION

### University of California, Berkeley — 2020 – Present

- Ph.D. in computer science advised by Stuart Russell and Anca Dragan, focusing on reinforcement learning theory, human-AI cooperation, and reward hacking.
- Supported by an Open Philanthropy AI Fellowship and a National Defense Science and Engineering Graduate (NDSEG) Fellowship.

### University of Maryland, College Park — 2014 – 2018

- Double B.S. in computer science and math, Middle East studies minor, GPA 4.00/4.
- Selected as one of top five graduating seniors for "academic distinction, exemplary character, and service to the campus or public communities."

## PUBLICATIONS AND PREPRINTS

**Cassidy Laidlaw**, Shivam Singhal, and Anca Dragan. **Correlated Proxies: A New Definition and Improved Mitigation for Reward Hacking**. *ICLR 2025* ***(spotlight)***.

Yaowen Ye*, **Cassidy Laidlaw***, and Jacob Steinhardt. **Iterative Label Refinement Matters More than Preference Optimization under Weak Supervision**. *ICLR 2025* ***(spotlight)***.

**Cassidy Laidlaw**, Banghua Zhu, Stuart Russell, and Anca Dragan. **The Effective Horizon Explains Deep RL Performance in Stochastic Environments**. *ICLR 2024* ***(spotlight)***.

Anand Siththaranjan*, **Cassidy Laidlaw***, and Dylan Hadfield-Menell. **Distributional Preference Learning: Understanding and Accounting for Hidden Context in RLHF**. *ICLR 2024*.

**Cassidy Laidlaw**, Stuart Russell, and Anca Dragan. **Bridging RL Theory and Practice with the Effective Horizon**. *NeurIPS 2023* ***(oral)***.

**Cassidy Laidlaw** and Anca Dragan. **The Boltzmann Policy Distribution: Accounting for Systematic Suboptimality in Human Models**. *ICLR 2022*.

**Cassidy Laidlaw** and Stuart Russell. **Uncertain Decisions Facilitate Better Preference Learning**. *NeurIPS 2021* ***(spotlight)***.

**Cassidy Laidlaw**, Sahil Singla, and Soheil Feizi. **Perceptual Adversarial Robustness: Defense Against Unseen Threat Models**. *ICLR 2021*.

**Cassidy Laidlaw** and Soheil Feizi. **Functional Adversarial Attacks**. *NeurIPS 2019*.

Daniel Cudeiro, Timo Bolkart, **Cassidy Laidlaw**, Anurag Ranjan, and Michael J. Black. **Capture, Learning, and Synthesis of 3D Speaking Styles**. *CVPR 2019*.

**Cassidy Laidlaw** and Soheil Feizi. **Playing it Safe: Adversarial Robustness with an Abstain Option**. *arXiv preprint (2019)*.

## EXPERIENCE

### Post-Undergraduate Researcher, University of Maryland — April 2019 – September 2020

- Research with Soheil Feizi on adversarial attacks and defenses for nonstandard threat models, leading to papers at NeurIPS 2019 and ICLR 2021.

**Research Intern,** Max Planck Institute for Intelligent Systems **May 2017 – January 2018**

- Research with Michael Black on statistical face models, leading to a paper at CVPR 2019.

**Freelance Software Developer** **June 2014 – August 2020**

- Built web, mobile, and data science solutions for startups, large corporations, and government.

## SERVICE AND OUTREACH

**Reviewing:** NeurIPS (2021-2024), ICLR (2022-2025), ICML (2023-2025), and various workshops. Received **outstanding reviewer award (top 8%)** for NeurIPS 2021.

**AI4ALL Project Leader:** led a group of high school students through a three-day AI project during this 2021 summer camp. Returned in 2022 to teach Python at the same camp.

## HONORS AND AWARDS

**Open Philanthropy AI Fellowship**

**National Defense Science and Engineering Graduate (NDSEG) Fellowship**

**Best Paper** at the 2023 ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems.

**Best Paper Honorable Mention** at the 2023 NeurIPS Workshop on Instruction Tuning and Instruction Following.

**University of Maryland University Medal Finalist:** selected as one of five finalists for the highest honor that the university can bestow on an undergraduate student based on the criteria of "academic distinction, exemplary character, and service to the campus or public communities."

**Banneker/Key Scholarship:** the University of Maryland's most prestigious scholarship.