RACHEL FREEDMAN

rachel.freedman@berkeley.edu | github.com/RachelFreedman | https://rachelfreedman.github.io/

Education

2019-2025 (expected) **Artificial Intelligence PhD, *UC Berkeley*** **3.93 GPA**

I work with the Center for Human-Compatible Artificial Intelligence (CHAI) on reinforcement learning, reward modeling, and interpretability. My advisor is Professor Stuart Russell.

2013-2017 Computer Science/Psychology BA, ***Duke University*** **3.93 GPA**

- Graduated Phi Beta Kappa and *magna cum laude*
- Earned High Distinction in Computer Science for my research thesis (published in the journal *Artificial Intelligence*)
- Designed original interdepartmental major entitled "Artificial Intelligence Systems" to explore interdisciplinary perspectives on AI

2015-2016 CS/Philosophy Registered Visiting Student, ***Oxford University*** **3.93 GPA**

Founded artificial intelligence student society

2011-2017 Part-Time Student, ***UNC Chapel Hill*** **4.00 GPA**

Publications

Rachel Freedman, Justin Svegliato, Kyle Wray, Stuart Russell. "Active Teacher Selection for Reward Learning". *Currently under review at ICLR 2023.*

Peter Barnett, **Rachel Freedman**, Justin Svegliato, Stuart Russell. "Active Reward Learning from Multiple Teachers". In *SafeAI at AAAI 2023*.
Best Paper Award Finalist at AAAI 2023 SafeAI Workshop.

Stephen Casper, [...], **Rachel Freedman**, [...], Dylan Hadfield-Menell. "Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback". *ArXiv Preprint.*

Oliver Daniels-Koch, **Rachel Freedman**. "The Expertise Problem: Learning from Specialized Feedback". In *ML Safety Workshop at NeurIPS* 2022.
AI Risk Analysis Award at NeurIPS 2022 ML Safety Workshop.

Rachel Freedman, Jana Schaich Borg, Walter Sinnott-Armstrong, John Dickerson, and Vincent Conitzer. "Adapting a Kidney Exchange Algorithm to Align with Human Values". *Artificial Intelligence*, v. 283, 2020. Also presented at *AAAI 2018, AIES 2018*, *MD4SG 2018*, and the *Participatory ML workshop* at *ICML 2020*.
Outstanding Student Paper Honorable Mention at *AAAI 2018*.

Rachel Freedman, Rohin Shah, and Anca Dragan. "Choice Set Misspecification in Reward Inference". In *Workshop on Artificial Intelligence Safety at IJCAI* 2020.
Best Paper Award at IJCAI 2020 AISafety Workshop.

Rohin Shah, Pedro Freire, Neel Alex, **Rachel Freedman**, Dmitrii Krasheninnikov, Lawrence Chan, Michael Dennis, Pieter Abbeel, Anca Dragan and Stuart Russell. "Benefits of Assistance over Reward Learning". In *Cooperative AI Workshop at NeurIPS* 2020.
Best Paper Award at NeurIPS 2020 CoopAI Workshop.

Work and Research

2019-pres. **Berkeley PhD Researcher** *CHAI lab (UC Berkeley)*

- Research topics include reinforcement learning, reward modeling, interpretability and LLM capability evaluation in support of robustly beneficial AI
- Contributed technical feedback to reports for the United Nations and the World Economic Forum and to Brian Christian's book *The Alignment Problem*
- Advised by Prof. Stuart Russell, also worked with Prof, Anca Dragan

2023 **Cambridge Visiting Researcher** *Krueger lab (Cambridge University)*

- Collaborated with Prof. David Krueger, Computational and Biological Learning Lab
- Researched reducing causal confusion in reward learning from human feedback

2017-2019 **Software Engineer and Consultant** *Galatea Associates (London, UK)*

- Designed and developed complex position-keeping and regulatory compliance solutions for financial institutions
- Collaborated with international team of consultants, engineers, contractors and stakeholders to manage ever-changing business problems and constraints
- Learned new tools and languages rapidly as required

2016-2017 **Moral AI Researcher** *Moral AI lab (Duke University)*

- Conducted a year-long independent research project incorporating computer science and psychology methodologies into the field of artificial morality
- Awarded High Distinction for my thesis research; presented and published this work at AAAI 2018, where it won *Outstanding Student Paper Honorable Mention*
- Published in *Artificial Intelligence*, poster presented at AIES 2018 and MD4SG 2018.

2015 summer **Software Engineering Intern** *Microsoft (Seattle, US)*

Honors

Conferences & Workshops

Best Paper Award Finalist *SafeAI Workshop at AAAI (2023)*
Invited Panelist *Growth Factory Venture Conference (2023)*
AI Risk Analysis Award *ML Safety Workshop at NeurIPS (2022)*
Invited Speaker *Institute for Advanced Study WAM Program (2022)*
Ambassador *Effective Altruism Global (2020 - 2021)*
Best Paper Award *CoopAI Workshop at NeurIPS (2020)*
Best Paper Award *AISafety Workshop at IJCAI-PRICAI (2020)*
Outstanding Student Paper Honorable Mention *AAAI (2018)*

Graduate

Foresight Fellow *Foresight Institute (2024)*
Manifund Grant *Manifund (2023)*
Rising Star in AI Ethics *Women in AI Ethics (2021)*
EECS Excellence Award *UC Berkeley (2019)*
EECS Departmental Fellowship *UC Berkeley (2019)*

Undergraduate

Phi Beta Kappa *Duke University (2017)*
magna cum laude *Duke University (2017)*
High Distinction in Computer Science *Duke University (2017)*
Robertson Scholarship *Robertson Scholarship Leadership Program (2013-2017)*
Thomas J. Watson Memorial Scholarship *IBM (2013-2016)*

Other

Rhodes Scholarship Finalist *Rhodes Trust (2017)*
Gates Cambridge Finalist *Gates Cambridge Trust (2019)*
President's Volunteer Service Gold Award *US Government (2013)*

Service and Leadership

2022-pres **AI Safety Advisor** *80000 Hours (contractor)*

2022 **Workshop Reviewer** *NeurIPS ML Safety Workshop (MLSW22)*

2022 **Journal Reviewer** *Journal of Artificial Intelligence (JAIR)*

2021 **Program Committee** *IJCAI Workshop on Artificial Intelligence Safety (NYC, US)*

2019-2021 **Ambassador** *Effective Altruism Global Conferences (London, UK and San Francisco, US)*

2015-2016 **Co-founder and President** *Oxford Existential Risk Society (Oxford, UK)*

Invited Talks and Panels

2023	**Panel on AI and Business** *Growth Factory Venture Conference*
2023	**Active Teacher Selection** *Cambridge University Krueger Lab*
2022	**Approaches to AI Safety** *EAGx Berkeley*
2022	**Value Alignment** *Institute for Advanced Study WAM program*
2022	**My Approach to Alignment Research** *SERI MATS seminar series*
2022	**Reward Modeling and Human Model Misspecification** *CHAI Workshop*
2022	**Panel on AI** *Duke University Career Center*
2021	**Panel on Graduate School** *AI Safety Support*

Advising and Mentoring

2023-ongoing	**Henry Papadatos**	EPFL MSc	
2022-2023	**Peter Barnett**	CHAI Intern	AAAI Workshop **(Award Finalist)**
2022-2023	**Oliver Daniels-Koch**	CHAI Intern	NeurIPS Workshop **(Award)**

Skills and Hobbies

Programming Python, NumPy, PyTorch, Java, Julia, Git; extensive practice learning new technologies quickly and proactively via engineering and consulting work

Service Mentor young people seeking a career in AI (particularly those from underrepresented backgrounds) on an ongoing basis (4+ years), volunteer at the Berkeley public animal shelter (1+ years), tutored and mentored at-risk North Carolina students weekly (2 years, earned the President's Volunteer Service Gold Award), volunteered full-time to support students in rural Mississippi (3 months).