

RACHEL FREEDMAN

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

Education

2019-2025 (expected)	Artificial Intelligence PhD, UC Berkeley I work with the Center for Human-Compatible Artificial Intelligence (CHAI) on AI safety via reinforcement learning, reward modeling, and assistance games. My advisor is Professor Stuart Russell.	3.93 GPA
2013-2017	Computer Science/Psychology BA, Duke University <ul style="list-style-type: none">• Graduated Phi Beta Kappa and <i>magna cum laude</i>• Earned High Distinction in Computer Science for my research thesis (now published in the journal <i>Artificial Intelligence</i>)• Designed original interdepartmental major entitled "Artificial Intelligence Systems" to explore interdisciplinary perspectives on AI	3.93 GPA
2015-2016	CS/Philosophy Registered Visiting Student, Oxford University Co-founded "Stop the Clock", an existential risk research discussion society that held weekly meetings at the Future of Humanity Institute	3.93 GPA
2011-2017	Part-Time Student, UNC Chapel Hill	4.00 GPA

Publications

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.

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 Krashenninikov, 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.	CHAI PhD Researcher UC Berkeley (Berkeley, CA) <ul style="list-style-type: none">• Research topics include reinforcement learning, active learning, and reward modeling in support of robustly beneficial and safe AI• Pivoted from interdisciplinary undergraduate to technical graduate program via extensive independent study of foundational mathematics, reinforcement learning, deep learning and advanced robotics alongside coursework and research• Contributed technical feedback to reports for the United Nations and the World Economic Forum and to Brian Christian's book <i>The Alignment Problem</i>• Advised by Prof. Stuart Russell, also worked with Prof. Anca Dragan
------------	---

2017-2019	Software Engineer and Consultant <i>Galatea Associates (London, UK)</i> <ul style="list-style-type: none"> 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 <i>Duke University (Durham, US)</i> <ul style="list-style-type: none"> Conducted a year-long independent research project incorporating computer science and psychology methodologies into the field of artificial morality Adapted a kidney exchange algorithm to align with human values by modeling underlying frameworks in human responses to moral dilemmas 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 <i>Artificial Intelligence</i>, poster presented at AIES 2018 and MD4SG 2018.
2015 summer	Software Engineering Intern <i>Microsoft (Seattle, US)</i>
2013-2014	Psychology and Neuroscience Research Assistant <i>Duke University (Durham NC, US)</i>

Honors

Conferences & Workshops	Best Paper Award Finalist <i>SafeAI Workshop at AAAI (2023)</i> AI Risk Analysis Award <i>ML Safety Workshop at NeurIPS (2022)</i> Invited Talk <i>Institute for Advanced Study WAM Program (2022)</i> Ambassador <i>Effective Altruism Global (2020 - 2021)</i> Best Paper Award <i>CoopAI Workshop at NeurIPS (2020)</i> Best Paper Award <i>AI Safety Workshop at IJCAI-PRICAI (2020)</i> Outstanding Student Paper Honorable Mention <i>AAAI (2018)</i>
Graduate	Rising Star in AI Ethics <i>Women in AI Ethics (2021)</i> EECS Excellence Award <i>UC Berkeley (2019)</i> EECS Departmental Fellowship <i>UC Berkeley (2019)</i>
Undergraduate	Phi Beta Kappa <i>Duke University (2017)</i> magna cum laude <i>Duke University (2017)</i> High Distinction in Computer Science <i>Duke University (2017)</i> Robertson Scholarship <i>Robertson Scholarship Leadership Program (2013-2017)</i> Thomas J. Watson Memorial Scholarship <i>IBM (2013-2016)</i>
Other	Rhodes Scholarship Finalist <i>Rhodes Trust (2017)</i> Gates Cambridge Finalist <i>Gates Cambridge Trust (2019)</i> President's Volunteer Service Gold Award <i>US Government (2013)</i>

Service and Leadership

2022-pres	AI Safety Advisor <i>80000 Hours (contractor)</i>
2022	Intern Advisor <i>Center for Human-Compatible AI (Berkeley, CA)</i>
2022	Workshop Reviewer <i>NeurIPS ML Safety Workshop (MLSW22)</i>
2022	Journal Reviewer <i>Journal of Artificial Intelligence (JAIR)</i>
2021	Program Committee <i>IJCAI Workshop on Artificial Intelligence Safety (NYC, US)</i>
2019-2021	Ambassador <i>Effective Altruism Global Conferences (London, UK and San Francisco, US)</i>
2015-2016	Co-founder and President <i>Oxford Existential Risk Society (Oxford, UK)</i>

Invited Talks and Panels

2022	Approaches to AI Safety <i>EAGx Berkeley</i>
2022	Value Alignment <i>Institute for Advanced Study WAM program</i>
2022	My Approach to Alignment Research <i>SERI MATS seminar series</i>
2022	Reward Modeling and Human Model Misspecification <i>CHAI Workshop</i>
2022	Panel on AI <i>Duke University Career Center</i>
2021	Panel on Graduate School <i>AI Safety Support</i>

Advising and Mentoring

2022-2023	Peter Barnett	Intern	AAAI Workshop (Award Finalist)
2022-2023	Oliver Daniels-Koch	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	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), volunteer at the public Berkeley animal shelter (1+ years), mentor young people seeking a career in AI (particularly those from underrepresented backgrounds) on an ongoing basis (3+ years)
Martial Arts	Earned World Taekwondo Federation-certified master (4th degree) black belt in Taekwondo after a decade of training. Trained, taught and competed for 12 years.
Creative Writing	Served as writer-in-residence at Italian artist residency "Arte Studio Ginestrelle." Stories and plays published at Duke, UNC and Oxford University student publications, and performed at Duke New Works Festival in 2014.