

# Adam Gleave

## EDUCATION

---

**University of California, Berkeley**, PhD in Artificial Intelligence. 2017–

My research focuses on developing techniques for advanced automated systems to act according to human preferences. I am supervised by Prof. Stuart Russell.

**University of Cambridge**, MPhil in Advanced Computer Science. 2015–2016

Graduated with **distinction** and awarded **Best Student Prize**, ranking 1<sup>st</sup> out of 31 students.

**University of Cambridge**, BA (Hons) in Computer Science. 2012–2015

Graduated with **first class** degree. Awarded **Best Student Prize** in 2014, ranking 1<sup>st</sup> out of 80 students, and in other years achieved a result in the top 10%.

## PUBLICATIONS

---

**Adam Gleave**, Michael Dennis, Shane Legg, Stuart Russell and Jan Leike.

“Quantifying Differences in Reward Functions”. In *International Conference on Learning Representations*, 2021. **Spotlight Paper** (top 20% of accepted papers).

Pedro Freire, **Adam Gleave**, Sam Toyer, Stuart Russell. “DERAIL: Diagnostic Environments for Reward and Imitation Learning”. In *DeepRL Workshop at NeurIPS*, 2020.

Eric J. Michaud, **Adam Gleave**, Stuart Russell. “Understanding Learned Reward Functions”. In *DeepRL Workshop at NeurIPS*, 2020.

**Adam Gleave**, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine and Stuart Russell. “Adversarial Policies: Attacking Deep Reinforcement Learning”. In *International Conference on Learning Representations*, 2020.

Aaron Tucker, **Adam Gleave** and Stuart Russell. “Inverse reinforcement learning for video games”. In *DeepRL Workshop at NeurIPS*, 2018.

**Adam Gleave** and Oliver Habryka. “Multi-task Maximum Causal Entropy Inverse Reinforcement Learning”. In *GoalsRL Workshop at ICML/IJCAI/AAMAS*, 2018.

Sören Mindermann, Rohin Shah, **Adam Gleave**, Dylan Hadfield-Menell. “Active Inverse Reward Design”. In *GoalsRL Workshop at ICML/IJCAI/AAMAS*, 2018.

**Adam Gleave** and Christian Steinruecken. “Making compression algorithms for Unicode text”. In *Data Compression Conference*, 2017.

Ionel Gog, Malte Schwarzkopf, **Adam Gleave**, Robert Watson and Steven Hand. “Firmament: fast, centralized cluster scheduling at scale”. In *Operating Systems Development And Implementation*, 2016.

## PROFESSIONAL & RESEARCH EXPERIENCE

---

**Research Intern**, DeepMind. May 2019–Oct 2019

DeepMind is an AI research lab. I worked with Jan Leike to develop a new method for evaluating reward models, leading to the publication “Quantifying Differences in Reward Functions” in ICLR 2021.

**Junior Researcher**, GSA Capital. October 2016–August 2017

GSA Capital is a quantitative hedge fund. I invented a futures trading strategy that was profitable in backtest and is now used in production. Extensive data analysis in Python with some development in Java and Scala.

**Trading Intern**, Jane Street Capital. June–September 2015

Jane Street is a proprietary trading firm. Created novel trading strategy for commodity desk which was profitable in out-of-sample data. Developed a model for the fixed income desk to assist pricing bond ETFs, now used in production.

**Developer Intern**, Jane Street Capital. June–September 2014

Optimized OCaml feed processor yielding  $50\times$  speedup; developed load testing framework leading to  $12\times$  performance improvement in internal protocol.

**Summer Intern**, Raspberry Pi. June–August 2013

Software engineering in C and Python for TAHMO: a low-cost meteorological station.

**Mathematics Intern**, i2OWater. August 2012

Devised a non-parametric model of pressure loss in water utility networks.

**Infrastructure Developer**, AquaMW. January–May 2012

Freelance Python software engineering for green-tech startup during high school.

## COMMUNITY CONTRIBUTIONS

---

**Open-source software.** Maintainer of [Stable Baselines](#) and [imitation](#), implementations of RL and imitation learning algorithms, with over 1500 stars on GitHub.

**Organizer** of NeurIPS 2019 Effective Altruism Social.

**Reviewer (Journals)** for JAIR, ACM Computing Surveys, IEEE Transactions on Artificial Intelligence.

**Reviewer (Conferences)** for ICLR 2021, NeurIPS 2020 (Top 10% reviewer), ICML 2020 (Top 33% Reviewer).

**Reviewer (Workshops)** for Cooperative AI (NeurIPS) 2020; SafeML Workshop (ICLR) 2019; I3 Workshop (ICML) 2019.

## SELECTED INVITED TALKS AND INTERVIEWS

---

[Understanding and Evaluating Learned Reward Functions.](#)

*BAIR Seminar, UC Berkeley.* March 2021.

[Adversarial Policies.](#)

*VITA Lab, EPFL.* December 2020.

[Forecasting AI Progress.](#)

*AI Impacts.* December 2019.

**Evaluating Reward Models.**

*DeepMind*. September 2019.

**Adversarial Policies: Attacking Deep Reinforcement Learning.**

*Future of Humanity Institute, University of Oxford*. June 2019.

**Scaling Inverse Reinforcement Learning for Human-Compatible AI.**

*WhiRL Lab, University of Oxford*. October 2018.

#### AWARDS

---

**Winton Capital Best MPhil Student Prize**, University of Cambridge. 2016

Awarded for the best result in the MPhil in Advanced Computer Science.

**College Scholarship**, St John's College, University of Cambridge. 2015

Scholarship providing full tuition and living costs, awarded on academic merit.

**Hockin (Wright) Prize**, St John's College, University of Cambridge. 2015

Prize for performance in third year Computer Science examinations.

**G-Research Best Student Prize**, University of Cambridge. 2014

Awarded for the best result in second year Computer Science examinations.

**Leathem (Wright) Prize**, St John's College, University of Cambridge. 2013

Prize for performance in first year Mathematics examinations.

**Pythagoras Prize**, St John's College, University of Cambridge. 2012

Full tuition scholarship, awarded to one student per year for mathematical aptitude.

#### TEACHING AND MENTORSHIP

---

**Teaching Assistant**, University of California, Berkeley. 2018-2019

*Introduction to AI*: presented weekly discussion sections, assisted students at office hours, graded exams and maintained website. *Safety and Control for AGI*: helped design curriculum for new course; designed coding project; delivered four lectures; grading.

**Career Mentor**, 80,000 Hours. 2018-2019

Providing career advice to early-stage students and professionals interested in pursuing AI research for social impact. Approximately 2 hours/week.

**Mathematics Instructor**, St John's College, University of Cambridge. 2015

Delivered intensive two-day course to eight incoming undergraduate mathematicians.