Adam Gleave

adam@gleave.me
gleave.me
AdamGleave

Experience

CEO & Co-Founder, FAR.AI. Jan 2022–present
FAR.AI is an AI safety research non-profit. As CEO, I set the overall agenda for our research projects and manage our technical staff to execute on this agenda.

Research Intern, DeepMind. Jan 2021–May 2021
DeepMind is an AI research lab. I worked with Dr Geoffrey Irving leading to the paper "Uncertainty Estimation for Language Reward Models".

Research Intern, DeepMind. May 2019–Oct 2019
I worked with Dr Jan Leike to develop a new method for evaluating reward models, leading to the paper "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 was deployed to production.

Trading Intern, Jane Street Capital. June–September 2015
Jane Street is a proprietary trading firm. Created novel commodity trading strategy profitable in out-of-sample data, and developed model now used by fixed income desk.

Developer Intern, Jane Street Capital. June–September 2014
Optimized OCaml feed processor yielding 50× speedup; developed load testing framework leading to 12× 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.

Education

University of California, Berkeley, PhD in Artificial Intelligence. 2017–2022
My thesis *Towards Trustworthy Machine Learning* focused on techniques for advanced automated systems to reliably act in accordance with human preferences. I was advised by Prof. Stuart Russell.

University of Cambridge, MPhil in Advanced Computer Science. 2015–2016
Graduated with **distinction**. Awarded **Best Student Prize** (1st 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 **1st** out of 80 students, and in other years achieved a result in the top 10%.

Publications

Joar Skalse, Lucy Farnik, Sumeet Ramesh Motwani, Erik Jenner, **Adam Gleave**, Alessandro Abate. "STARC: A General Framework For Quantifying Differences Between Reward Functions". In *International Conference on Learning Representations*, 2024.

Tony Tong Wang*, **Adam Gleave***, Tom Tseng, Kellin Pelrine, Nora Belrose, Joseph Miller, Michael D Dennis, Yawen Duan, Viktor Pogrebniak, Sergey Levine, Stuart Russell. "Adversarial Policies Beat Superhuman Go AIs". In *International Conference on Machine Learning*, 2023. **Oral Paper** (top 10% of accepted papers); and **Best Paper Award** at NeurIPS ML Safety Workshop 2022.

Joar Skalse, Matthew Farrugia-Roberts, Stuart Russell, Alessandro Abate, **Adam Gleave**. "Invariance in Policy Optimisation and Partial Identifiability in Reward Learning". In *International Conference on Machine Learning*, 2023.

Antonin Raffin, Ashley Hill, **Adam Gleave**, Anssi Kanervisto, Maximilian Ernestus, Noah Dormann. "Stable-Baselines3: Reliable Reinforcement Learning Implementations". In *Journal of Machine Learning Research*, 2021.

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).

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.

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.

Technical Reports and Workshop Papers

Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, **Adam Gleave**, Kellin Pelrine. "Scaling Laws for Data Poisoning in LLMs". In arXiv, 2024.

Nikolhaus Howe, Michał Zajac, Ian McKenzie, Oskar Hollinsworth, Tom Tseng, Pierre-Luc Bacon, **Adam Gleave**. "Exploring Scaling Trends in LLM Robustness". In *ICML NextGenAISafety Workshop*, 2024.

Adrià Garriga-Alonso, Mohammad Taufeeque, **Adam Gleave**. "Planning behavior in a recurrent neural network that plays Sokoban". In *ICML MI Workshop*, 2024.

Tom Tseng, Euan McLean, Kellin Pelrine, Tony T Wang, **Adam Gleave**. "Can Go AIs be adversarially robust?". In arXiv, 2024.

Pedro Freire, ChengCheng Tan, **Adam Gleave**, Dan Hendrycks, Scott Emmons. "Uncovering Latent Human Wellbeing in Language Model Embeddings". In arXiv, 2024.

Kellin Pelrine, Mohammad Taufeeque, Michał Zając, Euan McLean, **Adam Gleave**.

*Equal contribution.

"Exploiting novel GPT-4 APIs". In arXiv, 2023.

Lauro Langosco, David Krueger, **Adam Gleave**. "Training Equilibria in Reinforcement Learning". *NeurIPS Deep RL Workshop*, 2022.

Lev E McKinney, Yawen Duan, David Krueger, **Adam Gleave**. "On The Fragility of Learned Reward Functions". *NeurIPS ML Safety Workshop*, 2022.

Erik Jenner, Joar Max Viktor Skalse, **Adam Gleave**. "A general framework for reward function distances". *NeurIPS ML Safety Workshop*, 2022.

Adam Gleave*, Mohammad Taufeeque*, Juan Rocamonde*, Erik Jenner, Steven H. Wang, Sam Toyer, Maximilian Ernestus, Nora Belrose, Scott Emmons, Stuart Russell. "imitation: Clean Imitation Learning Implementations". In arXiv, 2022.

Erik Jenner, Herke Van Hoof, **Adam Gleave**. "Calculus on MDPs: Potential Shaping as a Gradient". In arXiv, 2022.

Pavel Czempin, **Adam Gleave**. "Reducing Exploitability with Population Based Training". In *New Frontiers in Adversarial Machine Learning at ICML*, 2022.

Adam Gleave*, Sam Toyer*. "A Primer on Maximum Causal Entropy Inverse Reinforcement Learning". In arXiv, 2022.

Adam Gleave, Geoffrey Irving. "Uncertainty Estimation for Language Reward Models". In arXiv, 2022.

Erik Jenner, **Adam Gleave**. "Preprocessing Reward Functions for Interpretability". In *NeurIPS Cooperative AI Workshop*, 2021.

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

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

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

Adam Gleave, 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.

Community Contributions

Open-source software. Maintainer of Stable Baselines, Stable Baselines3 and imitation, implementations of RL and imitation learning algorithms, with a total of over 10000 stars on GitHub.

Organizer of Center for Human Compatible AI Workshop 2023, 2022; Symposium on AGI Safety 2023, 2022; NeurIPS 2019 Human-Aligned AI Social.

Reviewer (Journals) for Artificial Intelligence (AIJ), Journal of Artificial Intelligence Research (JAIR), Journal of Machine Learning Research (JMLR), ACM Computing

Surveys, IEEE Transactions on Artificial Intelligence.
Reviewer (Conferences) for International Conference on Machine Learning (ICML) 2023, 2022 (Top 10% Reviewer); International Conference on Learning Representations (ICLR) 2023, 2022, 2021 (Top 33% Reviewer); Annual Conference on Neural Information Processing Systems (NeurIPS) 2022, 2021, 2020 (Top 10% reviewer)
Reviewer (Workshops) for Building and Evaluating Ethical Robotic Systems (IROS) 2021; Cooperative AI (NeurIPS) 2021, 2020; SafeML (ICLR) 2019; Imitation, Intent and Interaction (ICML) 2019.

Fund Manager for the Long-Term Future Fund between February 2020 and January 2022, working with of a team of 3-5 other fund managers to direct over $6 million in grant funding to effective charities working to improve humanity's long-term future.

Advising

Current Students
Pavel Czempin

Past Students
Joar Skalse (Ph.D. student in CS at Oxford), Matthew Farrugia-Roberts (MS student in CS at Melbourne), Lauro Langosco (Ph.D. student in CS at Cambridge), Oliver Richardson (Ph.D. student in CS at Cornell), Eric Michaud (Ph.D. student in Physics at MIT), Sergei Volodin, Pedro Freire (Research Engineer at FAR), Neel Kant (Research Staff at NVIDIA), Aaron Tucker (Ph.D. student in CS at Cornell), Erik Jenner (Ph.D. student in CS at Berkeley), Yawen Duan (MPhil. at Cambridge), Lev McKinney (Research Engineer at FAR AI), Leo Richter

Selected Press Coverage

Man beats machine at Go thanks to AI opponent's fatal flaw
The Times. February 2023.
Man beats machine at Go in human victory over AI
Financial Times. February 2023.
Scientists Found a Way to Defeat a 'Near-Superhuman' Go-Playing AI
VICE Motherboard. November 2022.
New Go-playing trick defeats world-class Go AI—but loses to human amateurs
Ars Technica. November 2022.
Can Your Career Help Change the World?
Wall Street Journal. October 2021.
Reinforcement-learning AIs are vulnerable to a new kind of attack
MIT Technology Review. February 2020.
Why deep-learning AIs are so easy to fool
Nature. October 2019.
The new high fliers giving it all away
The Telegraph, June 2015.

Selected Invited Talks and Interviews

AI Joins the Team: "Hello World!"
Gartner IT Symposium. Oct 2023.
Panel: The Future of Generalization: Scale, Safety and Beyond (Moderator)
ICML Workshop on Spurious Correlations, Invariance, and Stability. July 2023.
Adversarial Policies Beat Superhuman Go AIs
CHAI Workshop, Asilomar. June 2023.
The Tuned Lens
Mechanistic Interpretability Conference, MIT. May 2023.
Trustworthy Machine Learning
Guest Lecture in CS188, UC Berkeley. April 2023.
Inverse Reinforcement Learning
Guest Lecture in CS362, Stanford. April 2023.
Towards Trustworthy Machine Learning
SERI MATS. Jan 2023.
Robust Reward Learning
CHAI Workshop, Asilomar. June 2022.
Understanding and Evaluating Learned Reward Functions
BAIR Seminar, UC Berkeley. March 2021.
Computational and Biological Learning Lab, University of Cambridge. October 2021.
Seminar on Adversarial Policies
VITA Lab, EPFL. December 2020.
Podcast on Adversarial Policies
AXRP Podcast. 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

AI Fellowship Recipient, Open Philanthropy. 2021
Fellowship support for work improving the safety of transformative AI systems.

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.