

Scott Emmons

San Francisco, CA
www.scottemmons.com

Current Position

Google DeepMind,
Research Scientist.
AI Safety and Alignment.

June 2024 - Present

Education

UC Berkeley EECS,
PhD in Artificial Intelligence.
Advised by Stuart Russell.

2019 - 2024 (Expected)

University of North Carolina at Chapel Hill,
BS, Mathematics and BA, Computer Science.
Highest Honors for Thesis in Mathematics.

2015 - 2019

Honors & Awards

Department of Energy Computational Science Graduate Fellowship (\$300,000) *2019 - 2023*
Supports 4 years of graduate study for 20 U.S. students / year researching high-performance computing.

Churchill Scholarship (*declined*) *2019*
Awarded to 18 U.S. students / year to study for a master's degree at the University of Cambridge.

Robertson Scholars Leadership Program (\$250,000) *2015 - 2019*
Highly selective undergraduate merit scholarship providing dual citizenship at UNC and Duke.

Goldwater Scholarship (\$15,000) *2017 - 2019*
Awarded to 300 U.S. students / year for natural sciences, mathematics, and engineering research.

Archibald Henderson Medal *2019*
A gold medal, UNC's top undergraduate mathematics prize, given to 1 student / year.

Preprints

16. Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, **Scott Emmons**, Stuart Russell. "Evidence of Learned Look-Ahead in a Chess-Playing Neural Network." *arXiv*, 2024.
15. Leon Lang, Davis Foote, Stuart Russell, Anca Dragan, Erik Jenner, **Scott Emmons**. "When Your AIs Deceive You: Challenges of Partial Observability in Reinforcement Learning from Human Feedback." *arXiv*, 2024.
14. Alexandra Souly*, Qingyuan Lu*, Dillon Bowen*, Tu Trinh[†], Elvis Hsieh[†], Sana Pandey, Pieter Abbeel, Justin Svegliato, **Scott Emmons**, Olivia Watkins, Sam Toyer. "A Strong-REJECT for Empty Jailbreaks." *arXiv*, 2024.
13. Edmund Mills, Shiye Su, Stuart Russell, **Scott Emmons**. "ALMANACS: A Simulatability Benchmark for Language Model Explainability." *arXiv*, 2023.

Publications

12. Luke Bailey*, Euan Ong*, Stuart Russell, **Scott Emmons**. “Image Hijacks: Adversarial Images can Control Generative Models at Runtime.” *International Conference on Machine Learning (ICML)*, 2024.
11. Alexander Pan*, Jun Shern Chan*, Andy Zou*, Nathaniel Li, Steven Basart, Thomas Woodside, Hanlin Zhang, **Scott Emmons**, Dan Hendrycks. “Do the Rewards Justify the Means? Measuring Trade-Offs Between Rewards and Ethical Behavior in the Machiavelli Benchmark.” *International Conference on Machine Learning (ICML)*, 2023.
10. **Scott Emmons**, Caspar Oesterheld, Andrew Critch, Vincent Conitzer, Stuart Russell. “For Learning in Symmetric Teams, Local Optima are Global Nash Equilibria.” *International Conference on Machine Learning (ICML)*, 2022.
9. **Scott Emmons**, Benjamin Eysenbach, Ilya Kostrikov, Sergey Levine. “RvS: What is Essential for Offline RL via Supervised Learning?” *International Conference on Learning Representations (ICLR)*, 2022.
8. Xin Chen*, Sam Toyer*, Cody Wild*, **Scott Emmons**, Ian Fischer, Kuang-Huei Lee, Neel Alex, Steven H. Wang, Ping Luo, Stuart Russell, Pieter Abbeel, Rohin Shah. “An Empirical Investigation of Representation Learning for Imitation.” *Neural Information Processing Systems (NeurIPS)*, 2021.
7. **Scott Emmons***, Ajay Jain*, Michael Laskin*, Thanard Kurutach, Pieter Abbeel, Deepak Pathak. “Sparse Graphical Memory for Robust Planning.” *Neural Information Processing Systems (NeurIPS)*, 2020.
6. Eun Lee, **Scott Emmons**, Ryan Gibson, James Moody, Peter J. Mucha. “Concurrency and Reachability in Treelike Temporal Networks.” *Physical Review E*, 2019.
5. **Scott Emmons**, Peter J. Mucha. “Map Equation with Metadata: Varying the Role of Attributes in Community Detection.” *Physical Review E*, 2019.
4. Kris Hauser, **Scott Emmons**. “Global Redundancy Resolution via Continuous Pseudoinversion of the Forward Kinematic Map.” *IEEE Transactions on Automation Science and Engineering*, 2018.
3. **Scott Emmons**, Robert Light, Katy Börner. “MOOC Visual Analytics: Empowering Students, Teachers, Researchers, and Platform Developers of Massively Open Online Courses.” *Journal of the Association for Information Science and Technology (JASIST)*, 2017.
2. William H. Weir, **Scott Emmons**, Ryan Gibson, Dane Taylor, Peter J. Mucha. “Post-Processing Partitions to Identify Domains of Modularity Optimization.” *Algorithms*, 2017.
1. **Scott Emmons**, Stephen Kobourov, Mike Gallant, Katy Börner. “Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale.” *PLoS ONE*, 2016.

Software

- a. 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.” *arXiv*, 2022.

Leadership Experience

Center for Human-Compatible AI (CHAI), Berkeley, CA

Aug. 2019 - May 2024

PhD Student

- Co-managing CHAI's million-dollar compute budget by purchasing, installing, and maintaining an AI research cluster with 11 nodes, 88 GPUs, and 40 unique users.
- Co-managing CHAI's internship program, scaling it from 7 interns per year to 25 interns per year.

far.ai, Berkeley, CA

Feb. 2022 - July 2023

Cofounder and President

- Built FAR AI, Inc., a 501(c)(3) nonprofit that incubates and scales beneficial AI research agendas.
- Fundraised, recruited, and managed researchers to help define and execute on FAR's mission.

Invited Talks

Google DeepMind

April 18, 2024

Virtual.

“When Your AIs Deceive You: Challenges with Partial Observability of Human Evaluators in Reward Learning”

Technical AI Safety Conference 2024 (TAIS 2024)

April 5, 2024

Tokyo, Japan.

“When Your AIs Deceive You: Challenges with Partial Observability of Human Evaluators in Reward Learning”

Mentorship

Qingyuan Lu (Massachusetts Institute of Technology)

Leon Lang (University of Amsterdam)

Luke Bailey (Harvard University → Stanford University)

Edmund Mills (FAR AI → MultiOn)

Euan Ong (University of Cambridge)

Shiye Su (D. E. Shaw → University of Cambridge)

Michael Chen (Georgia Institute of Technology → Stripe)

Jiahai Feng (Massachusetts Institute of Technology → UC Berkeley)

Yulong Lin (University of Cambridge → Cohere)

Thomas Woodside (Yale University → Center for AI Safety)

Cynthia Chen (The University of Hong Kong → ETH Zurich)

Teaching

UC Berkeley's CS 188: Introduction to Artificial Intelligence

Graduate student instructor, spring 2022.

Outreach

Mentor for the Tianxia Fellowship, Center for Long Term Priorities, 2020.

Volunteer Service

Shanti Bhavan Children's Project, Tamil Nadu, India

July 2017 - Aug. 2017

Volunteer Teacher

- Taught approximately 80 primary and secondary school students from families who make less than \$2/day in subjects ranging from English literature to physics in preparation for employment and higher education.

Sunflower County Freedom Project, Sunflower, MS

May 2016 - July 2016

Volunteer Teacher

- Developed standard-aligned 8th- and 9th-grade math curriculum and taught it to two math classes that saw an average increase in performance of 9% on state standard test.

Professional Service

Conference Reviewing:

ICML 2021, 2022, 2024.

NeurIPS 2022, 2024.

ML Safety Workshop 2022.

Workshop Organization:

CHAI workshop 2024 program committee member.

PhD Fellowship Program Reviewer:

Future of Life Institute, Vitalik Buterin PhD Fellowship in AI Existential Safety, 2021.

Graduate Admissions Reviewer:

UC Berkeley EECS PhD application reviewer for incoming classes of 2021, 2022, 2023.