Scott Emmons

San Francisco, CA
www.scottemmons.com

Industry Experience

Google DeepMind, *2024 - 2025*
Research Scientist.
AI Safety and Alignment.

Education

UC Berkeley EECS, *2019 - 2024*
PhD in Computer Science.
Dissertation: *The Alignment Problem Under Partial Observability.*
Advised by Stuart Russell.

University of North Carolina at Chapel Hill, *2015 - 2019*
BS, Mathematics and BA, Computer Science.
Highest Honors for Thesis in Mathematics.

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

25. Max McGuinness*, Alex Serrano*, Luke Bailey, **Scott Emmons**. "Neural Chameleons: Language Models Can Learn to Hide Their Thoughts from Unseen Activation Monitors." *arXiv*, 2025.

24. **Scott Emmons***, Roland S. Zimmermann*, David K. Elson, Rohin Shah. "A Pragmatic Way to Measure Chain-of-Thought Monitorability." *arXiv*, 2025.

23. Tomek Korbak*, Mikita Balesni*, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, Mark Chen, Alan Cooney, Allan Dafoe, Anca Dragan, **Scott Emmons**, Owain Evans, David Farhi, Ryan Greenblatt, Dan Hendrycks, Marius Hobbhahn, Evan Hubinger, Geoffrey Irving, Erik Jenner, Daniel Kokotajlo, Victoria Krakovna, Shane Legg, David Lindner, David Luan, Aleksander Mądry, Julian Michael, Neel Nanda, Dave Orr, Jakub Pachocki, Ethan Perez, Mary Phuong, Fabien Roger, Joshua Saxe, Buck Shlegeris, Martín Soto, Eric Steinberger, Jasmine Wang, Wojciech Zaremba, Bowen Baker†, Rohin Shah†, Vlad Mikulik†. "Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety." *arXiv*, 2025.

22. **Scott Emmons**, Erik Jenner, David K. Elson, Rif A. Saurous, Senthooran Rajamanoharan, Heng Chen, Irhum Shafkat, Rohin Shah. "When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors." *arXiv*, 2025.

21. Rohin Shah, Alex Irpan*, Alexander Matt Turner*, Anna Wang*, Arthur Conmy*, David Lindner*, Jonah Brown-Cohen*, Lewis Ho*, Neel Nanda*, Raluca Ada Popa*, Rishub Jain*, Rory Greig*, Samuel Albanie*, **Scott Emmons***, Sebastian Farquhar*, Sébastien Krier*, Senthooran Rajamanoharan*, Sophie Bridgers*, Tobi Ijitoye*, Tom Everitt*, Victoria Krakovna*, Vikrant Varma*, Vladimir Mikulik*, Zachary Kenton*, Dave Orr*, Shane Legg, Noah Goodman, Allan Dafoe, Four Flynn, Anca Dragan. "An Approach to Technical AGI Safety and Security." *arXiv*, 2025.

20. Luke Bailey*, Alex Serrano*, Abhay Sheshadri*, Mikhail Seleznyov*, Jordan Taylor*, Erik Jenner*, Jacob Hilton, Stephen Casper, Carlos Guestrin, **Scott Emmons**. "Obfuscated Activations Bypass LLM Latent-Space Defenses." *arXiv*, 2024.

19. Edmund Mills, Shiye Su, Stuart Russell, **Scott Emmons**. "ALMANACS: A Simulatability Benchmark for Language Model Explainability." *arXiv*, 2023.

Publications

18. **Scott Emmons***, Caspar Oesterheld*, Vincent Conitzer, Stuart Russell. "Observation Interference in Partially Observable Assistance Games." *International Conference on Machine Learning (ICML)*, 2025.

17. Rylan Schaeffer, Dan Valentine, Luke Bailey, James Chua, Cristobal Eyzaguirre, Zane Durante, Joe Benton, Brando Miranda, Henry Sleight, Tony Tong Wang, John Hughes, Rajashree Agrawal, Mrinank Sharma, **Scott Emmons**, Sanmi Koyejo, Ethan Perez. "Failures to Find Transferable Image Jailbreaks Between Vision-Language Models." *International Conference on Learning Representations (ICLR)*, 2025.

16. Andrew Garber*, Rohan Subramani*, Linus Luu*, Mark Bedaywi, Stuart Russell, **Scott Emmons**. "The Partially Observable Off-Switch Game." *Association for the Advancement of Artificial Intelligence (AAAI)*, 2025.

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." *Neural Information Processing Systems (NeurIPS)*, 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 StrongREJECT for Empty Jailbreaks." *Neural Information Processing Systems (NeurIPS)*, 2024.

13. Erik Jenner, Shreyas Kapur, Vasil Georgiev, Cameron Allen, **Scott Emmons**, Stuart Russell. "Evidence of Learned Look-Ahead in a Chess-Playing Neural Network." *Neural Information Processing Systems (NeurIPS)*, 2024.

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

Center for Human-Compatible AI (CHAI), Berkeley, CA *Aug. 2019 - May 2024*
PhD Student

- Co-managed CHAI's **$1m compute budget** by purchasing, installing, and maintaining an AI research cluster with **11 nodes**, **88 GPUs**, and **40 users**.
- Co-managed CHAI's internship program, scaling it from 7 interns per year to 25 interns per year.
- Mentored **18 research interns** on projects resulting in **7 research papers**.

far.ai, Berkeley, CA *Feb. 2022 - July 2023*
Cofounder and President

- Fundraised and managed **$1.5m budget** of a 501(c)(3) AI safety research nonprofit.
- Hired **10 full-time employees and contractors** including a chief operations officer, seven research engineers, a special projects lead, and a communications specialist.
- Directed research project resulting in 1 research paper.

Teaching

UC Berkeley's CS 188: Introduction to Artificial Intelligence
Graduate student instructor, spring 2022.

Volunteer Service

Shanti Bhavan Children's Project, Tamil Nadu, India *July 2017 - Aug. 2017*
Volunteer Teacher

- Taught 80 primary and secondary school students from families making 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 8^{th}- and 9^{th}-grade math curriculum and taught it to two math classes that saw an average increase in performance of 9% on state standard test.

Invited Talks

OpenAI *July 21, 2025*
When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors

Center for Human-Compatible AI (Asilomar) *June 8, 2025*
Obfuscated Activations Bypass LLM Latent-Space Defenses

Singapore Alignment Workshop (Singapore) *April 23, 2025*
When Your AIs Deceive You: Challenges of Partial Observability in RLHF

Foresight Institute's Intelligent Cooperation Group *August 27, 2024*
When Your AIs Deceive You: Challenges of Partial Observability in RLHF

Center for Human-Compatible AI (Asilomar) *June 16, 2024*
When Your AIs Deceive You: Challenges of Partial Observability in RLHF

Google DeepMind *April 18, 2024*
When Your AIs Deceive You: Challenges of Partial Observability in RLHF

Technical AI Safety Conference (Tokyo, Japan) *April 5, 2024*
When Your AIs Deceive You: Challenges of Partial Observability in RLHF

United Kingdom AI Safety Institute *September 8, 2023*
Image Hijacks: Adversarial Images can Control Generative Models at Runtime

Department of Energy CSGF Program Review (Washington, D.C.) *July 19, 2023*
RvS: What is Essential for Offline RL via Supervised Learning?

Mentorship

Andrew Garber (Harvard University)
Linus Luu (University of Cambridge)
Mikhail Seleznyov (Skolkovo Institute of Science and Technology)
Alex Serrano Terré (Universitat Politècnica de Catalunya)
Rohan Subramani (Columbia University)
Mark Bedaywi (UC Berkeley)
Dillon Bowen (Ravio → FAR AI)
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 → Anthropic)
Shiye Su (D. E. Shaw → Stanford University)
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)

Outreach

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

Professional Service

Conference Reviewing:
ICML 2021, 2022, 2024.
NeurIPS 2022, 2024.
ML Safety Workshop 2022.

Workshop Organization:
Center for Human-Compatible AI 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.