SCOTT EMMONS

2121 BERKELEY WAY, OFFICE #8029, BERKELEY, CA 94720
EMAIL: SCOTT@SCOTTEMMONS.COM | WEBSITE: SCOTTEMMONS.COM

EDUCATION

University of California, Berkeley
PhD in Artificial Intelligence, *Department of Electrical Engineering and Computer Sciences*. 2019 - Present
Advised by Stuart Russell.

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

AWARDS AND HONORS

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

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

Goldwater Scholar ($15,000) 2017 - 2019
- Awarded to 300 students in the U.S. per 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 per year.

RESEARCH PREPRINTS

15. Leon Lang, Davis Foote, Stuart Russell, Anca Dragan, Erik Jenner, & **Scott Emmons**: "When Your AIs Deceive You: Challenges with Partial Observability of Human Evaluators in Reward Learning." *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 StrongREJECT for Empty Jailbreaks." *arXiv, 2024*.
13. Edmund Mills, Shiye Su, Stuart Russell, & **Scott Emmons**: "ALMANACS: A Simulatability Benchmark for Language Model Explainability." *arXiv, 2023*.
12. Luke Bailey[*], Euan Ong[*], Stuart Russell, & **Scott Emmons**: "Image Hijacks: Adversarial Images can Control Generative Models at Runtime." *arXiv, 2023*.

RESEARCH PUBLICATIONS

11. Alexander Pan[*], Chan Jun Shern[*], Andy Zou[*], Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, 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, Vince 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: "A 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**, Mike Gallant, Stephen Kobourov, & Katy Börner: "Analysis of Network Clustering Algorithms and Cluster Quality Metrics at Scale." *PLoS ONE, 2016*.

OPEN-SOURCE SOFTWARE

- 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
PhD Student August 2019 - Present

- 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
Cofounder and President February 2022 - July 2023

- 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.

SERVICE

Shanti Bhavan Children's Project Tamil Nadu, India
Volunteer Teacher July 2017 - August 2017

- 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
Volunteer Teacher May 2016 - July 2016

- 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.