ERIC J. MICHAUD

ericjm@mit.edu
eric.michaud99@gmail.com
ericjmichaud.com
github.com/ejmichaud

EDUCATION

Massachusetts Institute of Technology *2021-Present*
PhD student (Department of Physics)
Supervised by Max Tegmark

University of California, Berkeley *2017-2021*
Bachelor of Arts in Mathematics w/ Highest Honors
Minor in Computer Science
Highest Distinction in General Scholarship

PEER-REVIEWED PAPERS

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, ..., **Eric J Michaud**, ..., Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. Survey Certification

Eric J. Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The Quantization Model of Neural Scaling. *NeurIPS*, 2023

Ziming Liu, **Eric J. Michaud**, and Max Tegmark. Omnigrok: Gokking Beyond Algorithmic Data. *ICLR (Spotlight)*, 2023

Eric J. Michaud, Ziming Liu, and Max Tegmark. Precision Machine Learning. *Entropy*, 25(1), 2023

Ziming Liu, Ouail Kitouni, Niklas Nolte, **Eric J. Michaud**, Max Tegmark, and Mike Williams. Towards Understanding Grokking: An Effective Theory of Representation Learning. *NeurIPS (Oral)*, 2022

Scythia Marrow, **Eric J. Michaud**, and Erik Hoel. Examining the Causal Structures of Deep Neural Networks Using Information Theory. *Entropy*, 22(12):1429, 2020

Eric J. Michaud, Adam Gleave, and Stuart Russell. Understanding Learned Reward Functions. *Deep Reinforcement Learning Workshop at NeurIPS*, 2020

PREPRINTS AND WHITE PAPERS

Eric J. Michaud, Isaac Liao, Vedang Lad, Ziming Liu, Anish Mudide, Chloe Loughridge, Zifan Carl Guo, Tara Rezaei Kheirkhah, Mateja Vukelić, and Max Tegmark. Opening the AI black box: program synthesis via mechanistic interpretability. *arXiv preprint arXiv:2402.05110*, 2024

Eric J. Michaud, Andrew P. V. Siemion, Jamie Drew, and S. Pete Worden. Lunar Opportunities for SETI. *arXiv preprint arXiv:2009.12689*, 2020. A white paper for the National Academy of Sciences Planetary Science and Astrobiology Decadal Survey 2023-2032

Eric J. Michaud, Andrew P. V. Siemion, Jamie Drew, and S. Pete Worden. SETI from the lunar south pole. 2020. A white paper for the NASA Artemis III Science Definition Team

RESEARCH EXPERIENCE

Center for Human-Compatible AI May 2020 - August 2020
Research Intern

· Studied interpretability techniques for learned reward functions. Using a few different interpretability techniques, discovered interesting failure modes in some learned reward functions in simple RL environments.

Lawrence Livermore National Laboratory June 2019 - August 2019
DSSI Intern (Computational Engineering Division)

· Coded and ran physics simulations on LLNL clusters. Searched multi-dimensional experimental parameter space for notable behavior.
· Member of competitive Data Science Summer Institute program.

Berkeley SETI Research Center June 2018 - August 2018
Research Intern *(with periodic additional collaboration)*

· Worked on the idea of conducting SETI observations from the Moon's far side. Presented work at the 2019 International Astronautical Congress in Washington, D.C.
· Wrote networking software for the Breakthrough Listen Initiative computer cluster at the MeerKAT telescope.

TALKS

The Quantization Model of Neural Scaling, machine learning initiative seminar series, Perimeter Institute, Waterloo, Canada, October 2023

The Quantization Model of Neural Scaling, The 5th Workshop on Neural Scaling Laws: Emergence and Phase Transitions (Mila), July 2023

Omnigrok: Grokking Beyond Algorithmic Data, ICLR, May 2023, Kigali, Rwanda

The Quantization Model of Neural Scaling, Cengiz Pehlevan's reading group meeting, Harvard University, April 2023

The Quantization Model of Neural Scaling, David Bau's group meeting, Northeastern University, April 2023

The Quantization Model of Neural Scaling, SLT seminar, metauni, April 2023

Neural Scaling Exponents Beyond the Manifold Dimension, The 4th Workshop on Neural Scaling Laws (Mila), New Orleans, December 2022

SETI Space Telescope Mission Concepts, International Astronautical Congress, Paris, September 2022

Surprising capability gains in deep learning: grokking, 3rd Neural Scaling Laws Workshop, Mila Quebec, June 2022

Curious Properties of Neural Networks, IAIFI Journal Club, MIT, November 2021

Putting SETI in Space, and on the Moon, This Decade, Berkeley SETI Research Center, January 2021

Understanding Learned Reward Functions, Center for Human-Compatible AI, September 2020

Lunar Opportunities for SETI, International Astronautical Congress, Washington D.C., October 2019

WORK EXPERIENCE

MoonVu April 2017 - October 2017
Co-Founder, Engineer

· Built software for detecting, classifying, and simultaneously tracking objects in video. Advised on early company strategy. Left the company to focus on school work.

AWARDS

NSF Graduate Research Fellowship
National Merit Finalist
Dean's List

ACTIVITIES

The Berkeley Forum February 2019 - December 2019
Programming and Events Committees

· The Berkeley Forum organizes public talks and debates on the UC Berkeley campus. As a member of the programming and events committees, I organized public talks by Dr. Jill Tarter and by Jaron Lanier, as well as a debate on whether Big Tech should be broken up.