Eric J. Michaud

Department of Physics, MIT
ericjm [at] mit.edu
eric.michaud99 [at] gmail.com
`https://ericjmichaud.com`

Education

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

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

Experience

Center for Human-Compatible AI *Summer 2020*
Research Intern
Tested various techniques for interpreting and discovering failure modes in learned reward functions. The resulting paper, *Understanding Leared Reward Functions*, was accepted at the Deep RL Workshop at NeurIPS 2020.
Project code at: `https://github.com/HumanCompatibleAI/interpreting-rewards`

Lawrence Livermore National Laboratory *Summer 2019*
Intern (Computational Engineering Division)
Coded and ran physics simulations on national lab computer clusters. Searched multi-dimensional experimental parameter space of a physical system (in simulation) for notable behavior.
Member of competitive Data Science Summer Institute program.

Berkeley SETI Research Center *Summer 2018, misc 2019*
Intern (with periodic additional collaboration)
Wrote networking software for the Breakthrough Listen Initiative computer cluster at the MeerKAT telescope. Later, 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.

Publications

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

Joshua Engels, Isaac Liao, **Eric J. Michaud**, Wes Gurnee, and Max Tegmark. Not all language model features are linear. *arXiv preprint arXiv:2405.14860*, 2024

Xiaoman Delores Ding, Zifan Carl Guo, **Eric J. Michaud**, Ziming Liu, and Max Tegmark. Survival of the fittest representation: A case study with modular addition. *arXiv preprint arXiv:2405.17420*, 2024

Samuel Marks, Can Rager, **Eric J. Michaud**, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *arXiv preprint arXiv:2403.19647*, 2024

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

Talks

2023 Oct Perimeter Institute, Machine Learning Initiative seminar series, Waterloo, Canada
2023 Jul Mila Fifth Workshop on Neural Scaling Laws: Emergence and Phase Transitions
2023 May ICLR Spotlight talk, Kigali, Rwanda: *Omnigrok: Grokking Beyond Algorithmic Data*
2023 Apr Cengiz Pehlevan reading group, Harvard University, Cambridge, MA
2023 Apr David Bau group meeting, Northeastern University, Boston, MA
2023 Apr Singular Learning Theory seminar, metauni
2022 Dec Mila Fourth Workshop on Neural Scaling Laws, New Orleans, LA
2022 Sep International Astronautical Congress, IAA Symposium on SETI, Paris, France
2022 Jun Mila Third Neural Scaling Laws Workshop, Quebec, Canada
2021 Nov IAIFI Journal Club, MIT, (virtual)
2021 Jan Berkeley SETI Research Center (virtual)
2020 Sep Center for Human-Compatible AI (virtual)
2019 Oct International Astronautical Congress, IAA Symposium on SETI, Washington, D.C.

Awards

NSF Graduate Research Fellowship
National Merit Finalist
Dean's List

Other

MoonVu *Apr 2017 - Oct 2017*
Co-Founder, Engineer
Wrote software for performing multi-object detection and tracking, with the hope of counting the number of Tesla vehicles being shipped from their Fremont factory during the Model 3 production ramp up. The goal was to sell this count as alternative financial data. But my system wasn't accurate enough, and I got bogged down with school work, so quit.

The Berkeley Forum *Feb 2019 - Dec 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 invited Jill Tarter and Jaron Lanier and organized their talks. I also helped organize a debate on breaking up Big Tech.