Eric J. Michaud

eric.michaud99@gmail.com
https://ericjmichaud.com

Education

Massachusetts Institute of Technology *2021 - 2025*
PhD in Physics
Supervised by Max Tegmark
Thesis: *Decomposing Deep Neural Network Minds into Parts*

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

Experience

Astera Institute / Simplex AI Safety *Jan 2026 - Present*
Research Scientist

Goodfire AI *Summer 2025*
Resident
Studied how feature manifolds impact sparse autoencoder scaling behavior.

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 Learned 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

Eric J. Michaud, Liv Gorton, and Tom McGrath. Understanding sparse autoencoder scaling in the presence of feature manifolds. *Mechanistic Interpretability Workshop at NeurIPS*, 2025

Eric J. Michaud, Asher Parker-Sartori, and Max Tegmark. On the creation of narrow ai: hierarchy and nonlocality of neural network skills. *NeurIPS*, 2025

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, **Eric J. Michaud**, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability. *TMLR*, 2025

Anish Mudide, Joshua Engels, **Eric J. Michaud**, Max Tegmark, and Christian Schroeder de Witt. Efficient dictionary learning with switch sparse autoencoders. *ICLR*, 2025

Yuxiao Li*, **Eric J. Michaud***, David D Baek*, Joshua Engels, Xiaoqing Sun, and Max Tegmark. The geometry of concepts: Sparse autoencoder feature structure. *Entropy*, 27(4):344, 2024

Joshua Engels, **Eric J. Michaud**, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. *ICLR*, 2025

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. *ICLR (Oral)*, 2025

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: Distilling machine-learned algorithms into code. *Entropy*, 26(12):1046, 2024

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, Tony Wang, Samuel Marks, Charbel-Raphaël Segerie, Micah Carroll, Andi Peng, Phillip Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, **Eric J. Michaud**, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Bıyık, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open problems and fundamental limitations of reinforcement learning from human feedback. *TMLR*, 2023. Outstanding Paper Finalist

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

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

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

Jamie Simon, Daniel Kunin, Alexander Atanasov, Enric Boix-Adserà, Blake Bordelon, Jeremy Cohen, Nikhil Ghosh, Florentin Guth, Arthur Jacot, Mason Kamb, Dhruva Karkada, **Eric J. Michaud**, Berkan Ottlik, and Joseph Turnbull. There will be a scientific theory of deep learning. *arXiv preprint arXiv:2604.21691*, 2026

Ziming Liu, Yixuan Liu, **Eric J. Michaud**, Josh Gore, and Max Tegmark. Physics of skill learning. *arXiv preprint arXiv:2501.12391*, 2025

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

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

Talks

2026 May EPFL, FLARE Workshop, Lausanne, Switzerland
2025 Nov The Enigma Project, Stanford, CA
2025 Nov Simplex AI Safety / Astera Institute, Emeryville, CA
2025 Mar Redwood Center for Theoretical Neuroscience, Berkeley, CA
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