A Physics of Systems that Learn

Eric J. Michaud

January 2024 (revised October 2024)

Our world contains complexity and order at many levels. From fundamental particles, to atoms, to molecules, to systems of many particles, to biological life, nature follows physical laws. At each level, new abstractions and theories are needed to organize our understanding of nature [1]. The job of a physicist is to organize the apparent complexity of the world, at any level, under the most basic mathematical principles with the widest possible reach.

Much of the complexity we see in our everyday lives is the complexity of life. We are surrounded by organisms with intricate structure, structure produced by evolution over very long timescales. But there is another level of complexity to the world above this – a universe of things produced not over evolutionary time, but which are rather the result of what organisms *learn* during the course of their lives. Indeed, most of the complexity we experience at human scales in our environment is this sort of complexity – of ideas and artifacts which result from organisms like ourselves learning, thinking, planning, and acting to shape our environment. It is the complexity of intelligence.

How can we approach the problem of understanding this level of phenomena in the world? Of course, a great deal of effort already been put into understanding intelligent systems from neuroscientists, psychologists, computer scientists, and indeed from physicists [2]. But the present moment is an extraordinary time for studying learning and intelligence. Our moment is special because, with the success of deep learning, we are just now beginning to replicate increasingly general forms of intelligence artificially.

Observing this recent success, we are confronted with a couple of key facts: The first is that while the capabilities of today's best AI systems are still limited, they demonstrate that at least some aspects of intelligence are not constrained to biology and brains, but are rather substrate independent, a very general affordance of matter. Brains are no longer the only systems to study, if one wants to study general forms of intelligence. The second is that a new type of highly quantitative, empirical science on intelligent systems is now possible which wasn't before. With artificial neural networks, one can perform arbitrary interventions, experiments, and measurements, with access to the full internal state of the systems.

With this highly quantitative science now possible, physicists have flourished. Indeed, it is astonishing how important physicists have been so far in developing both theoretical and practical understanding of deep learning. To list some examples, there is the empirical work of Dario Amodei (PhD, Princeton, 2011), Jared Kaplan (PhD, Harvard, 2009), Sam McCandlish (PhD, Stanford, 2017), and Tom Henighan (PhD, Stanford, 2017) on neural scaling laws [3, 4], which inspired OpenAI to scale up autoregressive generative models to GPT-3 and beyond. There is also the theoretical work of Yasaman Bahri (PhD, UC Berkeley, 2017) [5], Dan Roberts (PhD, MIT, 2016) [6], and Cengiz Pehlevan (PhD, Brown, 2011) [7] and others on understanding these scaling laws theoretically. There is the work of Surya Ganguli (PhD, UC Berkeley, 2004) [8, 9], Hidenori Tanaka (PhD, Harvard, 2018) [10, 11], and many others on neural network loss landscapes and training dynamics. There is also the recent work of Adam Jermyn (PhD (Astronomy) Cambridge, 2018) and Adam Scherlis (Phd, Stanford, 2019) [12] on feature superposition in neural networks.

It seems then that there is already a field, populated in no small part by physicists, dedicated to understanding deep neural networks. Noticing this, it may make sense to formally recognize and define the beginnings of a subfield of physics here, dedicated to studying these systems, with the broader goal of understanding intelligence. It is worth responding immediately to some likely objections to this idea:

Isn't this just biophysics? While the study of intelligence and neural networks has traditionally been encompassed by biophysics, the field of understanding deep learning has some notable methodological differences from biophysics. Indeed, the background knowledge – both methods and terminology – are quite distinct from what biophysicists currently train in and typically encounter. To study AI systems, one needs knowledge of machine learning and all of its associated terminology and skills: knowledge of modern deep learning architectures and optimization methods and datasets and benchmarks – quite different from what is needed to study biological organisms, e.g. C. elegans.

Isn't this just computer science? Artificial intelligence has a long history within computer science. However, the mindset and methods typical of computer science are somewhat different from the mindset which has been most fruitful so far in understanding deep learning. As a discipline, computer science emerged out of mathematics [13], and is fairly attached to the methodology of mathematical proof. However, mathematical proofs have their limits in the insights they can generate about deep learning. Often, less formal (though still mathematical) models of phenomena in deep learning provide more insight. However a more basic limitation of mathematical formalism is that the properties of neural networks are often primarily determined by the properties of the data they are trained on. Absent a good *theory of data* then, there are limits to what one can derive from math alone.

It is worth spending slightly more time on this last point about data, since it gets at a core reason why physicists should be excited about studying deep neural networks, and more generally systems that learn. Consider what it means for a (ML) system to learn in the first place. It means that its internal state and dynamics in some way come to depend on facts about its (data) environment. In studying systems that learn, we can therefore hope to learn something new about the world. When we discover some order governing the properties of neural networks (e.g. scaling laws), these can hint at some corresponding order to the world, and to new physics waiting to be discovered about the organization of the world at a macro-level.

We have focused so far on the "science of deep learning", considered as a subfield of physics, with the broader goal of better understanding learning and intelligence and ultimately our world. However, there is another intersection between machine learning and physics which is a natural ally to this program: the application of machine learning as a tool directly for solving physics problems. In this case, the relationship between what ML models learn and new physical insights is more direct. The methods and terminology of each field are similar to each other, too. We might hope then that both machine learning applied to physics, and the physics of learning, could grow together, with more powerful models both directly and indirectly giving us greater insight into the organization of matter and information in our universe.

Acknowledgements: Many thanks to Isaac Chuang for encouraging me to write this and to William Brandon and Hidenori Tanaka for helpful discussions.

References

[1] Philip W Anderson. "More Is Different: Broken symmetry and the nature of the hierarchical structure of science." In: *Science* 177.4047 (1972), pp. 393–396.

[2] Hermann von Helmholtz. *Handbuch der physiologischen Optik*. Vol. 3. Third volume of a three-volume work. Leipzig: Leopold Voss, 1867.

[3] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. "Scaling laws for neural language models". In: *arXiv preprint arXiv:2001.08361* (2020).

[4] Tom Henighan, Jared Kaplan, Mor Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom B Brown, Prafulla Dhariwal, Scott Gray, Chris Hallacy, Benjamin Mann, Alec Radford, Aditya Ramesh, Nick Ryder, Daniel M Ziegler, John Schulman, Dario Amodei, and Sam McCandlish. "Scaling laws for autoregressive generative modeling". In: *arXiv preprint arXiv:2010.14701* (2020).

[5] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. "Explaining neural scaling laws". In: *arXiv preprint arXiv:2102.06701* (2021).

[6] Alexander Maloney, Daniel A Roberts, and James Sully. "A solvable model of neural scaling laws". In: *arXiv preprint arXiv:2210.16859* (2022).

[7] Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. "Spectrum dependent learning curves in kernel regression and wide neural networks". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1024–1034.

[8] Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. "Identifying and attacking the saddle point problem in high-dimensional non-convex optimization". In: *Advances in neural information processing systems* 27 (2014).

[9] Andrew M Saxe, James L McClelland, and Surya Ganguli. "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks". In: *arXiv preprint arXiv:1312.6120* (2013).

[10] Ekdeep Singh Lubana, Eric J Bigelow, Robert P Dick, David Krueger, and Hidenori Tanaka. "Mechanistic mode connectivity". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 22965–23004.

[11] Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel LK Yamins, and Hidenori Tanaka. "Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics". In: *arXiv preprint arXiv:2012.04728* (2020).

[12] Adam Scherlis, Kshitij Sachan, Adam S Jermyn, Joe Benton, and Buck Shlegeris. "Polysemanticity and capacity in neural networks". In: *arXiv preprint arXiv:2210.01892* (2022).

[13] Alan Mathison Turing. "On computable numbers, with an application to the Entscheidungsproblem". In: *Proceedings of the London Mathematical Society* 58 (Nov. 1936), pp. 230–265. DOI: 10.1112/plms/s2-42.1.230.