The unreasonable effectiveness of deep learning in artificial intelligence

Terrence J. Sejnowski[a,b,1]

[a]Computational Neurobiology Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037; and [b]Division of Biological Sciences, University of California San Diego, La Jolla, CA 92093

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved November 22, 2019 (received for review September 17, 2019)

Deep learning networks have been trained to recognize speech, caption photographs, and translate text between languages at high levels of performance. Although applications of deep learning networks to real-world problems have become ubiquitous, our understanding of why they are so effective is lacking. These empirical results should not be possible according to sample complexity in statistics and nonconvex optimization theory. However, paradoxes in the training and effectiveness of deep learning networks are being investigated and insights are being found in the geometry of high-dimensional spaces. A mathematical theory of deep learning would illuminate how they function, allow us to assess the strengths and weaknesses of different network architectures, and lead to major improvements. Deep learning has provided natural ways for humans to communicate with digital devices and is foundational for building artificial general intelligence. Deep learning was inspired by the architecture of the cerebral cortex and insights into autonomy and general intelligence may be found in other brain regions that are essential for planning and survival, but major breakthroughs will be needed to achieve these goals.

deep learning | artificial intelligence | neural networks

In 1884, Edwin Abbott wrote *Flatland: A Romance of Many Dimensions* (1) (Fig. 1). This book was written as a satire on Victorian society, but it has endured because of its exploration of how dimensionality can change our intuitions about space. Flatland was a 2-dimensional (2D) world inhabited by geometrical creatures. The mathematics of 2 dimensions was fully understood by these creatures, with circles being more perfect than triangles. In it a gentleman square has a dream about a sphere and wakes up to the possibility that his universe might be much larger than he or anyone in Flatland could imagine. He was not able to convince anyone that this was possible and in the end he was imprisoned.

We can easily imagine adding another spatial dimension when going from a 1-dimensional to a 2D world and from a 2D to a 3-dimensional (3D) world. Lines can intersect themselves in 2 dimensions and sheets can fold back onto themselves in 3 dimensions, but imagining how a 3D object can fold back on itself in a 4-dimensional space is a stretch that was achieved by Charles Howard Hinton in the 19th century (https://en.wikipedia.org/wiki/Charles_Howard_Hinton). What are the properties of spaces having even higher dimensions? What is it like to live in a space with 100 dimensions, or a million dimensions, or a space like our brain that has a million billion dimensions (the number of synapses between neurons)?

The first Neural Information Processing Systems (NeurIPS) Conference and Workshop took place at the Denver Tech Center in 1987 (Fig. 2). The 600 attendees were from a wide range of disciplines, including physics, neuroscience, psychology, statistics, electrical engineering, computer science, computer vision, speech recognition, and robotics, but they all had something in common: They all worked on intractably difficult problems that were not easily solved with traditional methods and they tended to be outliers in their home disciplines. In retrospect, 33 y later, these misfits were pushing the frontiers of their fields into high-dimensional spaces populated by big datasets, the world we are living in today. As the president of the foundation that organizes the annual NeurIPS conferences, I oversaw the remarkable evolution of a community that created modern machine learning. This conference has grown steadily and in 2019 attracted over 14,000 participants. Many intractable problems eventually became tractable, and today machine learning serves as a foundation for contemporary artificial intelligence (AI).

The early goals of machine learning were more modest than those of AI. Rather than aiming directly at general intelligence, machine learning started by attacking practical problems in perception, language, motor control, prediction, and inference using learning from data as the primary tool. In contrast, early attempts in AI were characterized by low-dimensional algorithms that were handcrafted. However, this approach only worked for well-controlled environments. For example, in blocks world all objects were rectangular solids, identically painted and in an environment with fixed lighting. These algorithms did not scale up to vision in the real world, where objects have complex shapes, a wide range of reflectances, and lighting conditions are uncontrolled. The real world is high-dimensional and there may not be any low-dimensional model that can be fit to it (2). Similar problems were encountered with early models of natural languages based on symbols and syntax, which ignored the complexities of semantics (3). Practical natural language applications became possible once the complexity of deep learning language models approached the complexity of the real world. Models of natural language with millions of parameters and trained with millions of labeled examples are now used routinely. Even larger deep learning language networks are in production today, providing services to millions of users online, less than a decade since they were introduced.

Origins of Deep Learning

I have written a book, *The Deep Learning Revolution: Artificial Intelligence Meets Human Intelligence* (4), which tells the story of how deep learning came about. Deep learning was inspired by the massively parallel architecture found in brains and its origins can be traced to Frank Rosenblatt's perceptron (5) in the 1950s that was based on a simplified model of a single neuron introduced by McCulloch and Pitts (6). The perceptron performed pattern recognition and learned to classify labeled examples (Fig. 3). Rosenblatt proved a theorem that if there was a set of parameters that could classify new inputs correctly, and there were

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "The Science of Deep Learning," held March 13–14, 2019, at the National Academy of Sciences in Washington, DC. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019 colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/science-of-deep-learning.

Author contributions: T.J.S. wrote the paper.

The author declares no competing interest.

This article is a PNAS Direct Submission.

[1]Email: terry@salk.edu.

First published January 28, 2020.

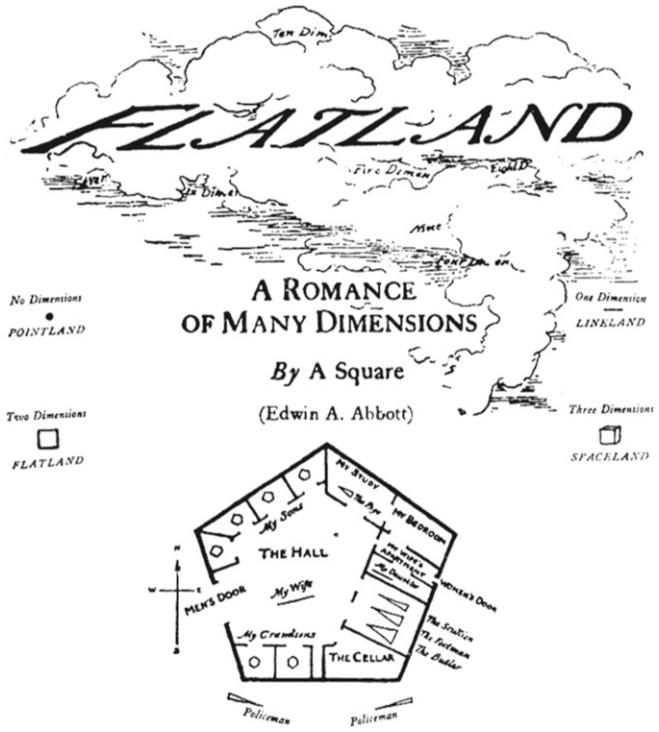

Fig. 1. Cover of the 1884 edition of *Flatland: A Romance in Many Dimensions* by Edwin A. Abbott (1). Inhabitants were 2D shapes, with their rank in society determined by the number of sides.

enough examples, his learning algorithm was guaranteed to find it. The learning algorithm used labeled data to make small changes to parameters, which were the weights on the inputs to a binary threshold unit, implementing gradient descent. This simple paradigm is at the core of much larger and more sophisticated neural network architectures today, but the jump from perceptrons to deep learning was not a smooth one. There are lessons to be learned from how this happened.

The perceptron learning algorithm required computing with real numbers, which digital computers performed inefficiently in the 1950s. Rosenblatt received a grant for the equivalent today of $1 million from the Office of Naval Research to build a large analog computer that could perform the weight updates in parallel using banks of motor-driven potentiometers representing variable weights (Fig. 3). The great expectations in the press (Fig. 3) were dashed by Minsky and Papert (7), who showed in their book *Perceptrons* that a perceptron can only represent categories that are linearly separable in weight space. Although at the end of their book Minsky and Papert considered the prospect of generalizing single- to multiple-layer perceptrons, one layer feeding into the next, they doubted there would ever be a way to train these more powerful multilayer perceptrons. Unfortunately, many took this doubt to be definitive, and the field was abandoned until a new generation of neural network researchers took a fresh look at the problem in the 1980s.

The computational power available for research in the 1960s was puny compared to what we have today; this favored programming rather than learning, and early progress with writing programs to solve toy problems looked encouraging. By the 1970s, learning had fallen out of favor, but by the 1980s digital computers had increased in speed, making it possible to simulate modestly sized neural networks. During the ensuing neural network revival in the 1980s, Geoffrey Hinton and I introduced a learning algorithm for Boltzmann machines proving that contrary to general belief it was possible to train multilayer networks (8). The Boltzmann machine learning algorithm is local and only depends on correlations between the inputs and outputs of single neurons, a form of Hebbian plasticity that is found in the cortex (9). Intriguingly, the correlations computed during training must be normalized by correlations that occur without inputs, which we called the sleep state, to prevent self-referential learning. It is also possible to learn the joint probability distributions of inputs without labels in an unsupervised learning mode. However, another learning algorithm introduced at around the same time based on the backpropagation of errors was much more efficient, though at the expense of locality (10). Both of these learning algorithm use stochastic gradient descent, an optimization technique that incrementally changes the parameter values to minimize a loss function. Typically this is done after averaging the gradients for a small batch of training examples.

Lost in Parameter Space

The network models in the 1980s rarely had more than one layer of hidden units between the inputs and outputs, but they were already highly overparameterized by the standards of statistical learning. Empirical studies uncovered a number of paradoxes that could not be explained at the time. Even though the networks were tiny by today's standards, they had orders of magnitude more parameters than traditional statistical models. According to bounds from theorems in statistics, generalization should not be possible with the relatively small training sets that were available. However,

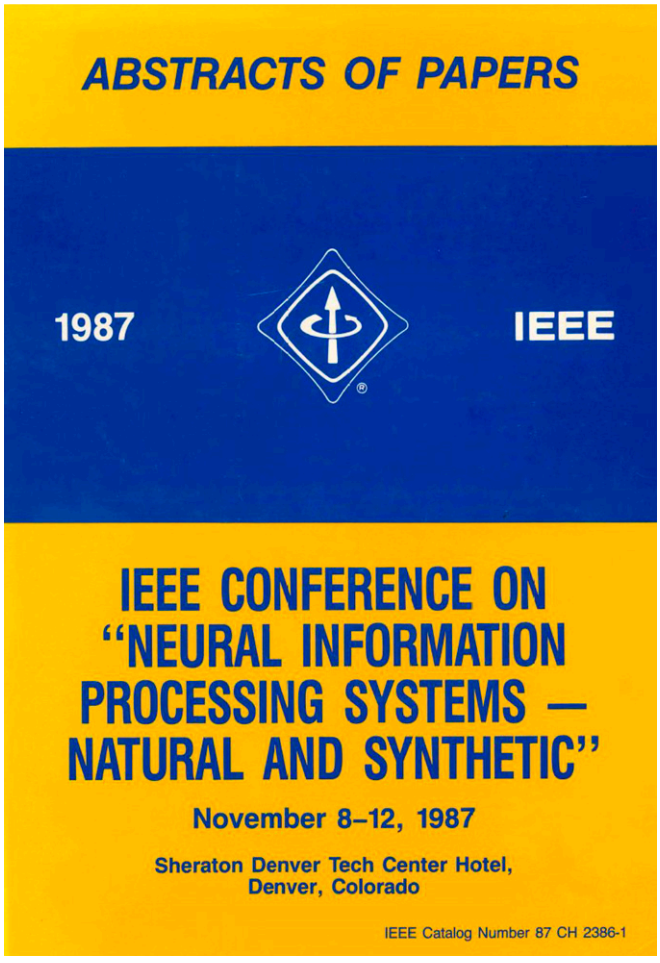

Fig. 2. The Neural Information Processing Systems conference brought together researchers from many fields of science and engineering. The first conference was held at the Denver Tech Center in 1987 and has been held annually since then. The first few meetings were sponsored by the IEEE Information Theory Society.

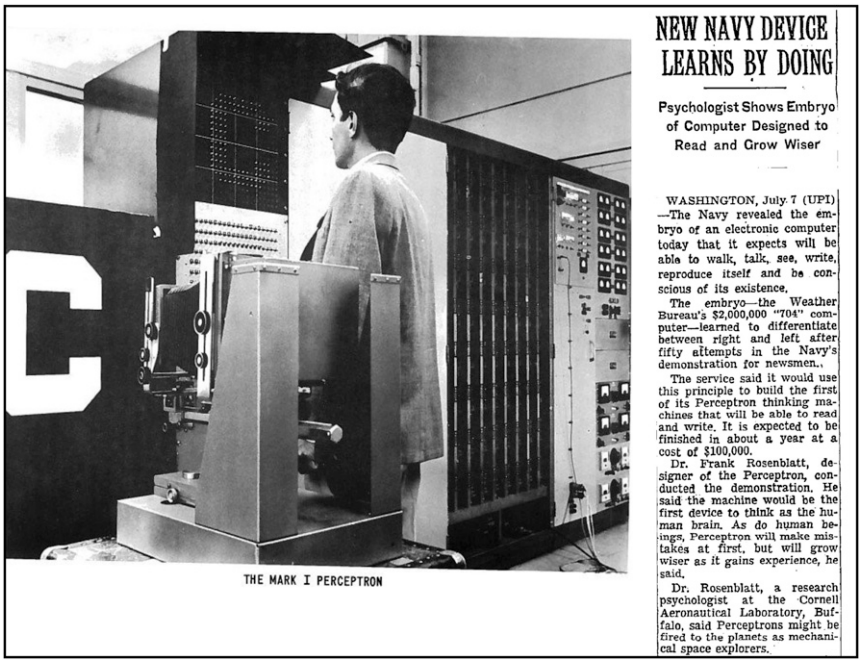

Fig. 3. Early perceptrons were large-scale analog systems (3). (*Left*) An analog perceptron computer receiving a visual input. The racks contained potentiometers driven by motors whose resistance was controlled by the perceptron learning algorithm. (*Right*) Article in the *New York Times*, July 8, 1958, from a UPI wire report. The perceptron machine was expected to cost $100,000 on completion in 1959, or around $1 million in today's dollars; the IBM 704 computer that cost $2 million in 1958, or $20 million in today's dollars, could perform 12,000 multiplies per second, which was blazingly fast at the time. The much less expensive Samsung Galaxy S6 phone, which can perform 34 billion operations per second, is more than a million times faster. Reprinted from ref. 5.

even simple methods for regularization, such as weight decay, led to models with surprisingly good generalization.

Even more surprising, stochastic gradient descent of nonconvex loss functions was rarely trapped in local minima. There were long plateaus on the way down when the error hardly changed, followed by sharp drops. Something about these network models and the geometry of their high-dimensional parameter spaces allowed them to navigate efficiently to solutions and achieve good generalization, contrary to the failures predicted by conventional intuition.

Network models are high-dimensional dynamical systems that learn how to map input spaces into output spaces. These functions have special mathematical properties that we are just beginning to understand. Local minima during learning are rare because in the high-dimensional parameter space most critical points are saddle points (11). Another reason why good solutions can be found so easily by stochastic gradient descent is that, unlike low-dimensional models where a unique solution is sought, different networks with good performance converge from random starting points in parameter space. Because of overparameterization (12), the degeneracy of solutions changes the nature of the problem from finding a needle in a haystack to a haystack of needles.

Many questions are left unanswered. Why is it possible to generalize from so few examples and so many parameters? Why is stochastic gradient descent so effective at finding useful functions compared to other optimization methods? How large is the set of all good solutions to a problem? Are good solutions related to each other in some way? What are the relationships between architectural features and inductive bias that can improve generalization? The answers to these questions will help us design better network architectures and more efficient learning algorithms.

What no one knew back in the 1980s was how well neural network learning algorithms would scale with the number of units and weights in the network. Unlike many AI algorithms that scale combinatorially, as deep learning networks expanded in size training scaled linearly with the number of parameters and performance continued to improve as more layers were added (13). Furthermore, the massively parallel architectures of deep learning networks can be efficiently implemented by multicore chips. The complexity of learning and inference with fully parallel hardware is O(1). This means that the time it takes to process an input is independent of the size of the network. This is a rare conjunction of favorable computational properties.

When a new class of functions is introduced, it takes generations to fully explore them. For example, when Joseph Fourier introduced Fourier series in 1807, he could not prove convergence and their status as functions was questioned. This did not stop engineers from using Fourier series to solve the heat equation and apply them to other practical problems. The study of this class of functions eventually led to deep insights into functional analysis, a jewel in the crown of mathematics.

The Nature of Deep Learning

The third wave of exploration into neural network architectures, unfolding today, has greatly expanded beyond its academic origins, following the first 2 waves spurred by perceptrons in the 1950s and multilayer neural networks in the 1980s. The press has rebranded deep learning as AI. What deep learning has done for AI is to ground it in the real world. The real world is analog, noisy, uncertain, and high-dimensional, which never jived with the black-and-white world of symbols and rules in traditional AI. Deep learning provides an interface between these 2 worlds. For example, natural language processing has traditionally been cast as a problem in symbol processing. However, end-to-end learning of language translation in recurrent neural networks extracts both syntactic and semantic information from sentences. Natural language applications often start not with symbols but with word embeddings in deep learning networks trained to predict the next word in a sentence (14), which are semantically deep and represent relationships between words as well as associations. Once regarded as "just statistics," deep recurrent networks are high-dimensional dynamical systems through which information flows much as electrical activity flows through brains.

One of the early tensions in AI research in the 1960s was its relationship to human intelligence. The engineering goal of AI was to reproduce the functional capabilities of human intelligence by writing programs based on intuition. I once asked Allen Newell, a computer scientist from Carnegie Mellon University and one of the pioneers of AI who attended the seminal Dartmouth summer conference in 1956, why AI pioneers had ignored brains, the substrate of human intelligence. The performance of brains was the only existence proof that any of the hard problems in AI could be solved. He told me that he personally had been open to insights from brain research but there simply had not been enough known about brains at the time to be of much help.

Over time, the attitude in AI had changed from "not enough is known" to "brains are not relevant." This view was commonly justified by an analogy with aviation: "If you want to build a flying machine, you would be wasting your time studying birds that flap their wings or the properties of their feathers." Quite to the contrary, the Wright Brothers were keen observers of gliding birds, which are highly efficient flyers (15). What they learned from birds was ideas for designing practical airfoils and basic principles of aerodynamics. Modern jets have even sprouted winglets at the tips of wings, which saves 5% on fuel and look suspiciously like wingtips on eagles (Fig. 4). Much more is now known about how brains process sensory information, accumulate evidence, make decisions, and plan future actions. Deep learning was similarly inspired by nature. There is a burgeoning new field in computer science, called algorithmic biology, which seeks to describe the wide range of problem-solving strategies used by biological systems (16). The lesson here is we can learn from nature general principles and specific solutions to complex problems, honed by evolution and passed down the chain of life to humans.

There is a stark contrast between the complexity of real neurons and the simplicity of the model neurons in neural network models. Neurons are themselves complex dynamical systems with a wide range of internal time scales. Much of the complexity of real neurons is inherited from cell biology—the need for each cell to generate its own energy and maintain homeostasis under a wide range of challenging conditions. However, other features of neurons are likely to be important for their computational function, some of which have not yet been exploited in model networks. These features include a diversity of cell types, optimized for specific functions; short-term synaptic plasticity, which can be either facilitating or depressing on a time scales of seconds; a cascade of biochemical reactions underlying plasticity inside synapses controlled by the history of inputs that extends from seconds to hours; sleep states during which a brain goes offline to restructure itself; and communication networks that control traffic between brain areas (17). Synergies between brains and AI may now be possible that could benefit both biology and engineering.

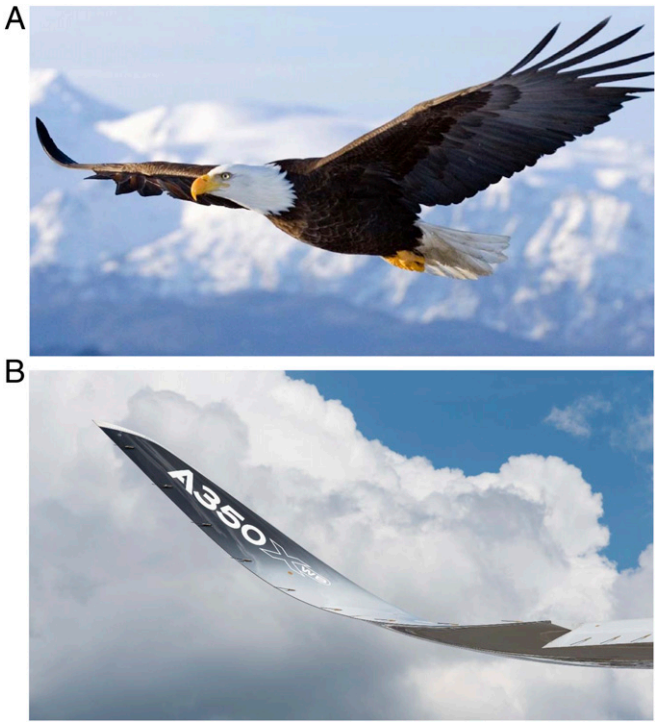

Fig. 4. Nature has optimized birds for energy efficiency. (*A*) The curved feathers at the wingtips of an eagle boosts energy efficiency during gliding. (*B*) Winglets on a commercial jets save fuel by reducing drag from vortices.

The neocortex appeared in mammals 200 million y ago. It is a folded sheet of neurons on the outer surface of the brain, called the gray matter, which in humans is about 30 cm in diameter and 5 mm thick when flattened. There are about 30 billion cortical neurons forming 6 layers that are highly interconnected with each other in a local stereotyped pattern. The cortex greatly expanded in size relative the central core of the brain during evolution, especially in humans, where it constitutes 80% of the brain volume. This expansion suggests that the cortical architecture is scalable—more is better—unlike most brain areas, which have not expanded relative to body size. Interestingly, there are many fewer long-range connections than local connections, which form the white matter of the cortex, but its volume scales as the 5/4 power of the gray matter volume and becomes larger than the volume of the gray matter in large brains (18). Scaling laws for brain structures can provide insights into important computational principles (19). Cortical architecture including cell types and their connectivity is similar throughout the cortex, with specialized regions for different cognitive systems. For example, the visual cortex has evolved specialized circuits for vision, which have been exploited in convolutional neural networks, the most successful deep learning architecture. Having evolved a general purpose learning architecture, the neocortex greatly enhances the performance of many special-purpose subcortical structures.

Brains have 11 orders of magnitude of spatially structured computing components (Fig. 5). At the level of synapses, each cubic millimeter of the cerebral cortex, about the size of a rice grain, contains a billion synapses. The largest deep learning networks today are reaching a billion weights. The cortex has the equivalent power of hundreds of thousands of deep learning networks, each specialized for solving specific problems. How are all these expert networks organized? The levels of investigation above the network level organize the flow of information between different cortical areas, a system-level communications problem. There is much to be learned about how to organize thousands of specialized networks by studying how the global flow of information in the cortex is managed. Long-range connections within the cortex are sparse because they are expensive, both because of the energy demand needed to send information over a long distance and also because they occupy a large volume of space. A switching network routes information between sensory and motor areas that can be rapidly reconfigured to meet ongoing cognitive demands (17).

Another major challenge for building the next generation of AI systems will be memory management for highly heterogeneous systems of deep learning specialist networks. There is need to flexibly update these networks without degrading already learned memories; this is the problem of maintaining stable, lifelong learning (20). There are ways to minimize memory loss and interference between subsystems. One way is to be selective about where to store new experiences. This occurs during sleep, when the cortex enters globally coherent patterns of electrical activity. Brief oscillatory events, known as sleep spindles, recur thousands of times during the night and are associated with the consolidation of memories. Spindles are triggered by the replay of recent

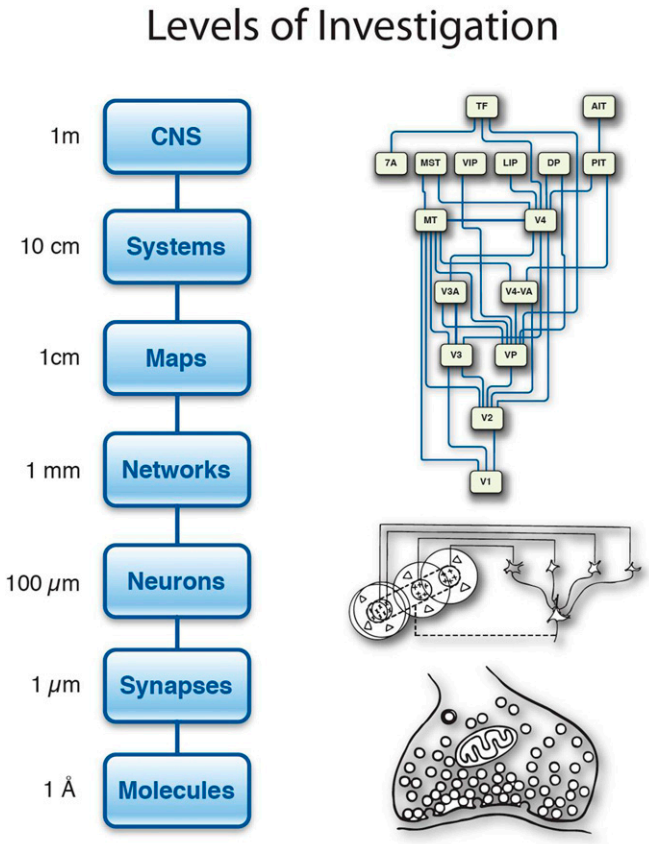

Fig. 5. Levels of investigation of brains. Energy efficiency is achieved by signaling with small numbers of molecules at synapses. Interconnects between neurons in the brain are 3D. Connectivity is high locally but relatively sparse between distant cortical areas. The organizing principle in the cortex is based on multiple maps of sensory and motor surfaces in a hierarchy. The cortex coordinates with many subcortical areas to form the central nervous system (CNS) that generates behavior.

episodes experienced during the day and are parsimoniously integrated into long-term cortical semantic memory (21, 22).

The Future of Deep Learning

Although the focus today on deep learning was inspired by the cerebral cortex, a much wider range of architectures is needed to control movements and vital functions. Subcortical parts of mammalian brains essential for survival can be found in all vertebrates, including the basal ganglia that are responsible for reinforcement learning and the cerebellum, which provides the brain with forward models of motor commands. Humans are hypersocial, with extensive cortical and subcortical neural circuits to support complex social interactions (23). These brain areas will provide inspiration to those who aim to build autonomous AI systems.

For example, the dopamine neurons in the brainstem compute reward prediction error, which is a key computation in the temporal difference learning algorithm in reinforcement learning and, in conjunction with deep learning, powered AlphaGo to beat Ke Jie, the world champion Go player in 2017 (24, 25). Recordings from dopamine neurons in the midbrain, which project diffusely throughout the cortex and basal ganglia, modulate synaptic plasticity and provide motivation for obtaining long-term rewards (26). Subsequent confirmation of the role of dopamine neurons in humans has led to a new field, neuroeconomics, whose goal is to better understand how humans make economic decisions (27). Several other neuromodulatory systems also control global brain states to guide behavior, representing negative rewards, surprise, confidence, and temporal discounting (28).

Motor systems are another area of AI where biologically inspired solutions may be helpful. Compare the fluid flow of animal movements to the rigid motions of most robots. The key difference is the exceptional flexibility exhibited in the control of high-dimensional musculature in all animals. Coordinated behavior in high-dimensional motor planning spaces is an active area of investigation in deep learning networks (29). There is also a need for a theory of distributed control to explain how the multiple layers of control in the spinal cord, brainstem, and forebrain are coordinated. Both brains and control systems have to deal with time delays in feedback loops, which can become unstable. The forward model of the body in the cerebellum provides a way to predict the sensory outcome of a motor command, and the sensory prediction errors are used to optimize open-loop control. For example, the vestibulo-ocular reflex (VOR) stabilizes image on the retina despite head movements by rapidly using head acceleration signals in an open loop; the gain of the VOR is adapted by slip signals from the retina, which the cerebellum uses to reduce the slip (30). Brains have additional constraints due to the limited bandwidth of sensory and motor nerves, but these can be overcome in layered control systems with components having a diversity of speed–accuracy trade-offs (31). A similar diversity is also present in engineered systems, allowing fast and accurate control despite having imperfect components (32).

Toward Artificial General Intelligence

Is there a path from the current state of the art in deep learning to artificial general intelligence? From the perspective of evolution, most animals can solve problems needed to survive in their niches, but general abstract reasoning emerged more recently in the human lineage. However, we are not very good at it and need long training to achieve the ability to reason logically. This is because we are using brain systems to simulate logical steps that have not been optimized for logic. Students in grade school work for years to master simple arithmetic, effectively emulating a digital computer with a 1-s clock. Nonetheless, reasoning in humans is proof of principle that it should be possible to evolve large-scale systems of deep learning networks for rational planning and decision making. However, a hybrid solution might also be possible, similar to neural Turing machines developed by DeepMind for learning how to copy, sort, and navigate (33). According to Orgel's Second Rule, nature is cleverer than we are, but improvements may still be possible.

Recent successes with supervised learning in deep networks have led to a proliferation of applications where large datasets are available. Language translation was greatly improved by training on large corpora of translated texts. However, there are many applications for which large sets of labeled data are not available. Humans commonly make subconscious predictions about outcomes in the physical world and are surprised by the unexpected. Self-supervised learning, in which the goal of learning is to predict the future output from other data streams, is a promising direction (34). Imitation learning is also a powerful way to learn important behaviors and gain knowledge about the world (35). Humans have many ways to learn and require a long period of development to achieve adult levels of performance.

Brains intelligently and spontaneously generate ideas and solutions to problems. When a subject is asked to lie quietly at rest in a brain scanner, activity switches from sensorimotor areas to a default mode network of areas that support inner thoughts, including unconscious activity. Generative neural network models can learn without supervision, with the goal of learning joint probability distributions from raw sensory data, which is abundant. The Boltzmann machine is an example of generative model (8). After a Boltzmann machine has been trained to classify inputs, clamping an output unit on generates a sequence of examples from that category on the input layer (36). Generative adversarial

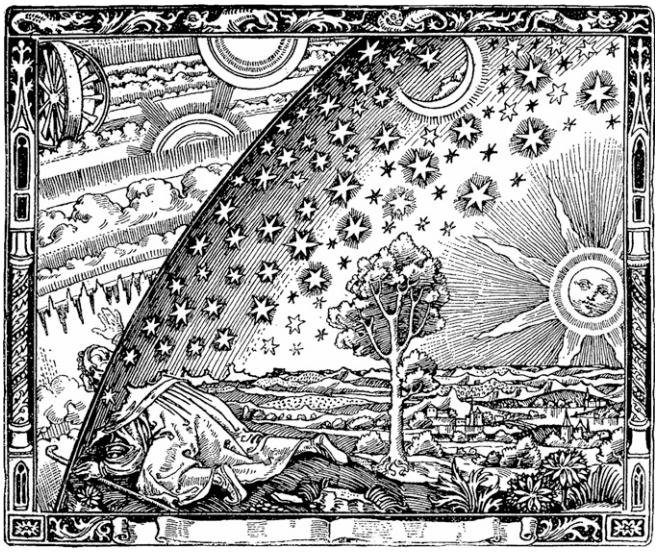

Fig. 6. The caption that accompanies the engraving in Flammarion's book reads: "A missionary of the Middle Ages tells that he had found the point where the sky and the Earth touch ...." Image courtesy of Wikimedia Commons/Camille Flammarion.

networks can also generate new samples from a probability distribution learned by self-supervised learning (37). Brains also generate vivid visual images during dream sleep that are often bizarre.

Looking ahead

We are at the beginning of a new era that could be called the age of information. Data are gushing from sensors, the sources for pipelines that turn data into information, information into knowledge, knowledge into understanding, and, if we are fortunate, knowledge into wisdom. We have taken our first steps toward dealing with complex high-dimensional problems in the real world; like a baby's, they are more stumble than stride, but what is important is that we are heading in the right direction. Deep learning networks are bridges between digital computers and the real world; this allows us to communicate with computers on our own terms. We already talk to smart speakers, which will become much smarter. Keyboards will become obsolete, taking their place in museums alongside typewriters. This makes the benefits of deep learning available to everyone.

In his essay "The Unreasonable Effectiveness of Mathematics in the Natural Sciences," Eugene Wigner marveled that the mathematical structure of a physical theory often reveals deep insights into that theory that lead to empirical predictions (38). Also remarkable is that there are so few parameters in the equations, called physical constants. The title of this article mirrors Wigner's. However, unlike the laws of physics, there is an abundance of parameters in deep learning networks and they are variable. We are just beginning to explore representation and optimization in very-high-dimensional spaces. Perhaps someday an analysis of the structure of deep learning networks will lead to theoretical predictions and reveal deep insights into the nature of intelligence. We can benefit from the blessings of dimensionality.

Having found one class of functions to describe the complexity of signals in the world, perhaps there are others. Perhaps there is a universe of massively parallel algorithms in high-dimensional spaces that we have not yet explored, which go beyond intuitions from the 3D world we inhabit and the 1-dimensional sequences of instructions in digital computers. Like the gentleman square in Flatland (Fig. 1) and the explorer in the Flammarion engraving (Fig. 6), we have glimpsed a new world stretching far beyond old horizons.

Data Availability. There are no data associated with this paper.

1. E. A. Abbott, *Flatland: A Romance in Many Dimensions* (Seeley & Co., London, 1884).
2. L. Breiman, Statistical modeling: The two cultures. *Stat. Sci.* **16**, 199–231 (2001).
3. N. Chomsky, *Knowledge of Language: Its Nature, Origins, and Use* (Convergence, Praeger, Westport, CT, 1986).
4. T. J. Sejnowski, *The Deep Learning Revolution: Artificial Intelligence Meets Human Intelligence* (MIT Press, Cambridge, MA, 2018).
5. F. Rosenblatt, *Perceptrons and the Theory of Brain Mechanics* (Cornell Aeronautical Lab Inc., Buffalo, NY, 1961), vol. VG-1196-G, p. 621.
6. W. S. McCulloch, W. H. Pitts, A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).
7. M. Minsky, S. Papert, *Perceptrons* (MIT Press, Cambridge, MA, 1969).
8. D. H. Ackley, G. E. Hinton, T. J. Sejnowski, A learning algorithm for Boltzmann Machines. *Cogn. Sci.* **9**, 147–169 (1985).
9. T. J. Sejnowski, The book of Hebb. *Neuron* **24**, 773–776 (1999).
10. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
11. R. Pascanu, Y. N. Dauphin, S. Ganguli, Y. Bengio, On the saddle point problem for non-convex optimization. arXiv:1405.4604 (19 May 2014).
12. P. L. Bartlett, P. M. Long, G. Lugosi, A. Tsigler, Benign overfitting in linear regression. arXiv:1906.11300 (26 June 2019).
13. T. Poggio, A. Banburski, Q. Liao, Theoretical issues in deep networks: Approximation, optimization and generalization. arXiv:1908.09375 (25 August 2019).
14. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality" in *Proceedings of the 26th International Conference on Neural Imaging Processing Systems* (Curran Associates, 2013), vol. 2, pp. 3111–3119.
15. D. McCullough, *The Wright Brothers* (Simon & Schuster, New York, 2015).
16. S. Navlakha, Z. Bar-Joseph, Algorithms in nature: The convergence of systems biology and computational thinking. *Mol. Syst. Biol.* **7**, 546 (2011).
17. S. B. Laughlin, T. J. Sejnowski, Communication in neuronal networks. *Science* **301**, 1870–1874 (2003).
18. K. Zhang, T. J. Sejnowski, A universal scaling law between gray matter and white matter of cerebral cortex. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5621–5626 (2000).
19. S. Srinivasan, C. F. Stevens, Scaling principles of distributed circuits. *Curr. Biol.* **29**, 2533–2540.e7 (2019).
20. G. Gary Anthes, Lifelong learning in artificial neural networks. *Commun. ACM* **62**, 13–15 (2019).
21. L. Muller *et al*., Rotating waves during human sleep spindles organize global patterns of activity during the night. *eLife* **5**, 17267 (2016).
22. R. Todorova, M. Zugaro, Isolated cortical computations during delta waves support memory consolidation. *Science* **366**, 377–381 (2019).
23. P. S. Churchland, *Conscience: The Origins of Moral Intuition* (W. W. Norton, New York, 2019).
24. Wikipedia, AlphaGo versus Ke Jie. https://en.wikipedia.org/wiki/AlphaGo_versus_Ke_Jie. Accessed 8 January 2020.
25. D. Silver *et al*., A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* **362**, 1140–1144 (2018).
26. P. R. Montague, P. Dayan, T. J. Sejnowski, A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
27. P. W. Glimcher, C. Camerer, R. A. Poldrack, E. Fehr, *Neuroeconomics: Decision Making and the Brain* (Academic Press, New York, 2008).
28. E. Marder, Neuromodulation of neuronal circuits: Back to the future. *Neuron* **76**, 1–11 (2012).
29. I. Akkaya *et al*., Solving Rubik's cube with a robot hand. arXiv:1910.07113 (16 October 2019).
30. S. du Lac, J. L. Raymond, T. J. Sejnowski, S. G. Lisberger, Learning and memory in the vestibulo-ocular reflex. *Annu. Rev. Neurosci.* **18**, 409–441 (1995).
31. Y. Nakahira, Q. Liu, T. J. Sejnowski, J. C. Doyle, Fitts' Law for speed-accuracy trade-off describes a diversity-enabled sweet spot in sensorimotor control. arXiv:1906.00905 (18 September 2019).
32. Y. Nakahira, Q. Liu, T. J. Sejnowski, J. C. Doyle, Diversity-enabled sweet spots in layered architectures and speed-accuracy trade-offs in sensorimotor control. arXiv:1909.08601 (18 September 2019).
33. A. Graves, G. Wayne, I. Danihelka, Neural turing machines. arXiv:1410.540 (20 October 2014).
34. A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, A. Torralba, Self-supervised audio-visual co-segmentation. arXiv:1904.09013 (18 April 2019).
35. S. Schaal, Is imitation learning the route to humanoid robots? *Trends Cogn. Sci.* **3**, 233–242 (1999).
36. G. E. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554 (2006).
37. I. J. Goodfellow *et al*., Generative adversarial nets. arXiv:1406.2661(10 June 2014).
38. E. P. Wigner, The unreasonable effectiveness of mathematics in the natural sciences. Richard Courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Commun. Pure Appl. Math.* **13**, 1–14 (1960).

人工知能における深層学習の理不尽な有効性

The unreasonable effectiveness of deep learning in artificial intelligence

Terrence J. Sejnowski

PNAS 2020, Vol. 117, no. 48, 30033-30038

概要

深層学習ネットワークは音声認識写真のキャプション言語間のテキスト翻訳などを高いレベルで実現している。深層学習ネットワークの実世界の問題への応用はいたるところで見られるようになった。だがなぜこれほど効果的なのかについての理解は不足している。これらの経験的な結果は統計学におけるサンプルの複雑さや非凸最適化理論によればあり得ないはずである。しかし深層学習ネットワークの学習と効果におけるパラドックスが調査されており、高次元空間の幾何学的な知見が得られている。深層学習の数学的理論は深層学習がどのように機能するかを明らかにしさまざまなネットワークアーキテクチャの長所と短所を評価できるようになり大きな改善につながるだろう。深層学習は人間がデジタル機器とコミュニケーションをとるための自然な方法を提供し、人工的な一般知能を構築するための基盤となっている。深層学習は大脳皮質のアーキテクチャから着想を得たものである。自律性や一般知能に関する知見は計画性や生存に不可欠な他の脳領域にある。だがこれらの目標を達成するには大きなブレークスルーが必要である。

1884 年エドウィン・アボットは「平坦な国: 多次元のロマンス (1)」を発表した (図1)。この本はヴィクトリア朝の社会を風刺するために書かれたものである。この絵は次元によって空間に対する私たちの直感がどのように変化するかを探求したことで、今でも語り継がれている。フラットランドは 2 次元の世界であり，幾何学的な生物が住んでいた。これらの生物は 2 次元の数学を完全に理解しており、円は三角形よりも完璧であった。その中で四角い紳士が球の夢を見て、自分の宇宙が自分やフラットランドの誰もが想像していたよりもずっと大きいかもしれないという可能性に目を覚ます。彼はそれが可能であることを誰にも納得させることができず、最後には投獄されてしまう。

1 次元の世界から 2 次元の世界へ、 2 次元の世界から 3 次元の世界へと、空間的な次元が増えることは容易に想像できる。 2 次元では線が交差し、3 次元ではシートが折り返される。 4 次元空間で 3 次元の物体が折り返されることを想像するのは 19 世紀にチャールズ・ハワード・ヒントンが実現したことである(https://en.wikipedia.org/wiki/Charles_Howard_Hinton)。さらに高い次元の空間にはどのような性質があるのか？ 100 次元や 100 万次元の空間、あるいは脳のように100億次元 (神経細胞間のシナプスの数) の空間では、どのような生活ができるのだろうか？

1987 年第 1 回 Neural Information Processing Systems (NeurIPS) Conference and Workshop がデンバー工業センターで開催された (図2)。 600 人の参加者は物理学, 神経科学, 心理学, 統計学, 電気工学, コンピュータサイエンス, コンピュータビジョン, 音声認識, ロボット工学など, さまざまな分野から集まっていた。だが彼らには共通点があった。彼らはいずれも従来の手法では容易に解決できない難解な問題に取り組んでいた。それぞれの分野で異彩を放っている傾向があった。 33 年後に振り返ってみるとこれらの異端児たちはそれぞれの分野のフロンティアを大規模なデータセットが存在する高次元空間へと押し広げていた。だがそれは現在の私たちが生きている世界である。私は毎年開催される NeurIPS 会議を主催する財団の理事長として、現代の機械学習を生み出したコミュニティの目覚ましい進化を見守ってきた。この会議は着実に成長し 2019 年には 14,000 人以上の参加者を集めた。多くの難解な問題がやがて扱いやすくなり、今日、機械学習は現代の人工知能 (AI) の基盤として機能している。

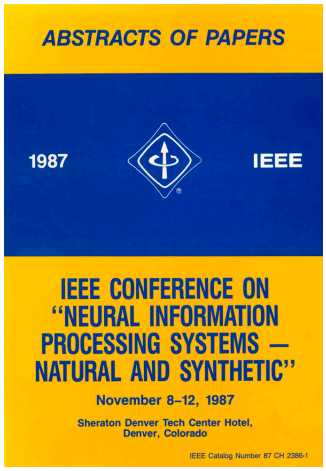

図2　Neural Information Processing Systems Conference には理工系のさまざまな分野の研究者が集まった。1987 年に Denver Tech Center で第 1 回会議が開催されて以来、毎年開催されている。最初の数回は IEEE Information Theory Society が主催していた。

機械学習の初期の目標は AI に比べてより控えめなものであった。機械学習は一般的な知能を直接目指すのではなくデータからの学習を主な手段として知覚, 言語, 運動制御, 予測, 推論などの実用的な問題に取り組むことから始まった。一方 AI の初期の試みは低次元のアルゴリズムを手作業で作ることが特徴であった。しかしこの方法は制御された環境でしか使えなかった。例えば積み木の世界では、すべての物体は長方形の固体で同じように塗装され固定された照明のある環境であった。しかしこれらのアルゴリズムは物体の形状が複雑で、反射率の範囲が広く、照明条件が制御できない実世界の視覚には対応できなかった。実世界は高次元でありそれに適合する低次元のモデルは存在しないかもしれない(2)。同様の問題は意味論の複雑さを無視した記号と構文に基づく初期の自然言語モデルにも見られた(3)。自然言語の実用化が可能になったのは深層学習言語モデルの複雑さが実世界の複雑さに近づいてからである。現在では数百万のパラメータを持ち数百万のラベル付き例を用いて学習された自然言語モデルが日常的に使用されている。さらに大規模な深層学習言語ネットワークは登場してから 10 年も経たないうちにオンラインで何百万人ものユーザーにサービスを提供するようになっている。

1　深層学習の起源

私は「The Deep Learning Revolution」という本を書きました(4)。この本には、深層学習がどのようにして生まれたかが書かれています。深層学習は脳の超並列アーキテクチャにヒントを得たもので、その起源は、1950 年代に McCulloch と Pitts が導入した単一ニューロンの単純化されたモデルをもとに、Frank Rosenblatt が開発したパーセプトロン (5) にあります(6)。パーセプトロンはパターン認識を行いラベルを付けた例を分類することを学習した(図3)。ローゼンブラットは新しい入力を正しく分類できるパラメータのセットが存在しかつ十分な数の例があれば、彼の学習アルゴリズムはそれを見つけることが保証されるという定理を証明した。学習アルゴリズムは、ラベル付きデータを使って、二値のしきい値ユニットへの入力の重みであるパラメータに小さな変更を加え、勾配降下法を実施した。しかしパーセプトロンから深層学習への移行は決してスムーズなものではありませんでした。しかしパーセプトロンから深層学習への移行は決してスムーズなものではありませんでした。その経緯には教訓があります。

パーセプトロンの学習アルゴリズムは 1950 年代のデジタルコンピューターでは非効率的な実数計算が必要だった。ローゼンブラットは 1950 年代に海軍研究局から 100 万ドルに相当する助成金を得てモーター駆動のポテンショメータで重みを変化させ、その重みの更新を並列に行うことができる大型のアナログコンピュータを製作した(図3)。記者による大きな期待 (図3) は、Minsky と Papert（7）によって打ち砕かれました。彼らは、著書『Perceptrons』の中でパーセプトロンは重み空間で線形分離可能なカテゴリーしか表現できないことを示しました。 Minsky と Papert は著書の最後で、単層パーセプトロンから多層パーセプトロンへの一般化を考えていましたが、彼らはより強力な多層パーセプトロンを学習する方法があるのかどうか疑問に思っていました。残念ながら、この疑念は決定的なものとして受け止められ、1980 年代に新しい世代のニューラルネットワーク研究者がこの問題を見直すまで、この分野は見捨てられていました。

1960 年代に研究に使用できた計算機の性能は現在のものに比べてちっぽけなものでした。そのため学習よりもプログラミングが好まれ、おもちゃの問題を解決するためにプログラムを書くことは、初期の段階では有望視されていました。1970 年代になると学習型の研究は下火になりました。だが 1980 年代になると、デジタルコンピュータが高速化し、適度な大きさのニューラルネットワークのシミュレーションが可能になりました。 1980 年代のニューラルネットワーク・リバイバルの中で Geoffrey Hinton と私はボルツマンマシン用の学習アルゴリズムを発表し、一般的な考えに反して多層ネットワークの学習が可能であることを証明しました(8)。ボルツマンマシンの学習アルゴリズムは局所的で、単一のニューロンの入力と出力の間の相関関係にのみ依存しています。興味深いことに、自己言及的な学習を防ぐために、学習中に計算された相関は、我々がスリープ状態と呼ぶ、入力なしで起こる相関によって正規化されなければならない。また、ラベルのない入力の結合確率分布を教師なし学習モードで学習す

ることも可能である。しかし、ほぼ同時期に登場した誤差のバックプロパゲーションに基づく別の学習アルゴリズムの方が、局所性を犠牲にしてもはるかに効率的であった(10)。これらの学習アルゴリズムはいずれも，損失関数を最小化するためにパラメータ値を段階的に変化させる最適化手法である確率的勾配降下法を用いています。これは，損失関数を最小化するためにパラメータの値を段階的に変更する最適化手法であり，通常，少量の学習例の勾配を平均化した後に実行される。

2　パラメータ空間の消失

1980 年代のネットワークモデルは入力と出力の間に 1 層以上の隠れユニットを持つことはほとんどなかった。だが統計的学習の基準からするとすでに高度にオーバーパラメータ化されていた。実証実験では当時としては説明のつかないパラドックスがいくつも発見された。ネットワークは現在の基準では小さなものであるにもかかわらず従来の統計モデルよりも桁違いに多くのパラメータを持っていた。統計学の定理に基づく境界線に従えば比較的小さなトレーニングセットでは一般化はできないはずである。しかし重み崩壊などの単純な正則化手法でも驚くほど優れた汎化能力を持つモデルが得られた。

さらに驚くべきことに非凸の損失関数の確率的勾配降下法では局所的な最小値に陥ることはほとんどなかった。誤差がほとんど変化しない下降途中の長いプラトーがありその後急激に下降する。これらのネットワークモデルとその高次元パラメータ空間の幾何学的性質のおかげで従来の直感で予測されていた失敗とは逆に解への効率的なナビゲートと優れた一般化を達成することができた。

ネットワークモデルは入力空間を出力空間にマッピングする方法を学習する高次元の動的システムである。これらの関数には特別な数学的特性があり我々はそれを理解し始めたところである。高次元のパラメータ空間ではほとんどの臨界点が鞍点であるため学習中に局所的な最小値をとることはほとんどない(11)。確率的勾配降下法で良い解が簡単に見つかるもう一つの理由はユニークな解が求められる低次元モデルとは異なり良い性能を持つ異なるネットワークはパラメータ空間のランダムな開始点から収束していくことである。過剰なパラメータ化 (12) により解が縮退することで問題の性質が「干し草の中の針を探す」から「干し草の中の針を探す」に変わる。

多くの疑問が残されている。なぜ少ない例と多くのパラメータから一般化することができるのか？なぜ確率的勾配降下法は他の最適化手法に比べて有用な関数を見つけるのに有効なのか？ある問題に対するすべての良い解の集合はどのくらいの大きさか？良い解はお互いに何らかの関係があるのか？一般化を向上させるためのアーキテクチャの特徴と帰納的バイアスの関係は？これらの質問に対する答えはより優れたネットワークアーキテクチャやより効率的な学習アルゴリズムの設計に役立つ。

1980 年代の時点ではニューラルネットワークの学習アルゴリズムが、ネットワークのユニットや重みの数に応じてどの程度拡張できるのか、誰も知らなかった。多くの AI アルゴリズムが組み合わせ的にスケーリングするのとは異なり、深層学習ネットワークのサイズが拡大すると、学習はパラメータの数に応じてリニアにスケーリングされ、層を増やすほどに性能が向上していきました(13)。さらに深層学習ネットワークの超並列アーキテクチャはマルチコアチップで効率的に実装することができる。完全に並列化されたハードウェアによる学習と推論の複雑さは O(1) です。これは入力を処理するのにかかる時間がネットワークのサイズに依存しないことを意味する。これは有利な計算特性の組み合わせとしては珍しいことである。

新しいクラスの関数が導入されると、それを完全に解明するには何世代もかかるものである。例えば 1807 年にフーリエがフーリエ級数を発表したとき収束を証明できず関数としての地位を疑問視された。しかし技術者たちはフーリエ級数を使って熱方程式を解くなど実用的な問題に応用していった。このような関数の研究は、最終的に数学の王冠の中の宝石ともいえる関数解析の深い洞察へとつながっていった。

3　深層学習の性質

1950 年代のパーセプトロン, 1980 年代の多層ニューラルネットワークに続く, ニューラルネットワークアーキテクチャの探求の第 3 の波は、学術的な起源を超えて大きく広がっている。マスコミは深層学習を AI と言い換えている。深層学習が AI にもたらしたものは現実の世界に根ざしたものである。現実の世界はアナログでノイズが多く不確実で高次元であり従来の AI の記号や規則といった白黒の世界とは相容れないものであった。この 2 つの世界をつなぐのが深層学習です。例えば自然言語処理は従来、記号処理の問題とされてきた。しかしリカレントニューラルネットワークにおける言語翻訳のエンドツーエンドの学習は文から構文と意味の両方の情報を抽出する。自然言語処理のアプリケーションは記号ではなく文中の次の単語を予測するように訓練された深層学習ネットワークの単語埋め込みから始まることが多い(14)。この単語埋め込みは意味的に深く単語間の関係だけでなく関連性も表している。かつては「単なる統計」と考えられていた深層再帰ネットワークは脳に電気的活動が流れるように情報が流れる高次元の動的システムである。

1960 年代の AI 研究における初期の緊張感のひとつは人間の知能との関係であった。 AI の工学的な目標は直感に基づいてプログラムを書くことで人間の知能の機能的な能力を再現することであった。カーネギーメロン大学のコンピュータ科学者で 1956 年のダートマス夏季会議に参加したAIの先駆者の一人であるアレン・ニューウェルに「なぜ AI の先駆者たちは人間の知能の基質である脳を無視したのか」と尋ねたことがある。脳の性能は AI の難問が解決できる唯一の存在証明であった。彼は自分自身は脳の研究から得られる洞察を受け入れていた。だが当時は脳について十分な知識がなかったためあまり役に立たなかったと言っていた。

時が経つにつれ AI の考え方は「十分なことはわかっていない」から「脳は関係ない」へと変わっていった。この考え方は「空飛ぶ機械を作りたければ、羽ばたく鳥やその羽の特性を研究しても時間の無駄だ」という航空学との類似性によって正当化されるのが一般的だった。それどころかライト兄弟は、飛行効率の高い滑空する鳥を熱心に観察していたのだ(15)。彼らが鳥から学んだのは実用的な翼型を設計するためのアイデアと空気力学の基本原理だった。最近のジェット機では、翼の先端にウィングレットを設けて燃料を 5 %節約している。だが，これはワシの翼の先端に酷似する(図4)。脳がどのように感覚情報を処理し、証拠を蓄積し、意思決定を行い、将来の行動を計画するかについては、多くのことが分かっています。深層学習も同じように自然からヒントを得ている。コンピュータサイエンスの分野ではアルゴリズム生物学と呼ばれる新しい分野が急成長しており、生物システムが使用する幅広い問題解決戦略を説明しようとしている(16)。ここでの教訓は、複雑な問題に対する一般的な原理と具体的な解決策を自然から学ぶことができるということである。これらは、進化によって磨かれ、生命の連鎖を経て人間に受け継がれてきた。

実際のニューロンの複雑さと、ニューラルネットワークモデルのモデルニューロンの単純さは対照的である。ニューロンは、それ自体がさまざまな時間スケールを持つ複雑な動的システムである。実際のニューロンの複雑さの多くは、細胞生物学から受け継いだものである。すなわち、各細胞が自らエネルギーを生成し、さまざまな厳しい条件の下でホメオスタシスを維持する必要がある。しかし、神経細胞の計算機能にとって重要な特徴は他にもあり、そのうちのいくつかはモデルネットワークではまだ利用されていない。これらの特徴には、特定の機能に最適化された多様な細胞タイプ、数秒の時間スケールで促進的にも抑制的にもなりうる短期的なシナプス可塑性などがある。数秒から数時間に及ぶ入力の履歴によって制御される、シナプス内部の可塑性を支える生化学反応のカスケード。脳をオフラインにして再構築する睡眠状態、脳領域間のトラフィックを制御する通信ネットワークなどがあります(17)。脳と AI の相乗効果は、生物学と工学の両方に恩恵をもたらす可能性があります。

新皮質は 2 億年前に哺乳類に出現した。灰白質と呼ばれる脳の外側にある神経細胞の折り畳まれたシートで、人間の場合、平らにすると直径約 30 cm 厚さ約 5 mmになる。大脳皮質の神経細胞は約 300 億個あり 6 つの層を形成しており、局所的な定型パターンで相互に強く結びついている。大脳皮質は、進化の過程で脳の中心部に比べて大きく拡大し、特にヒトでは脳容積の 80 %を占めている。この拡大は、体の大きさに比べて拡大していないほとんどの脳領域とは異なり、大脳皮質の構造はスケーラブルで、多ければ多いほど良いことを示唆している。興味深いことに、大脳皮質の白質を形成する局所的な結合に比べ、長距離的な結合は非常に少ないのですが、その体積は灰白質の 5/4 乗に比例し、大きな脳では灰白質の体積よりも大きくなる(18)。脳構造のスケーリング則は、重要な計算原理を知る上でのヒントになる(19)。細胞の種類とその結合性を含む大脳皮質の構造は、大脳皮質全体で類似しており、異なる認知システムに特化した領域が存在する。例えば、視覚野では、視覚に特化した回路が進化し、深層学習アーキテクチャとして最も成功している畳み込みニューラルネットワークに利用されている。大脳新皮質は、汎用の学習アーキテクチャを進化させたことで、多くの特殊目的の皮質下構造の性能を大きく向上させている。

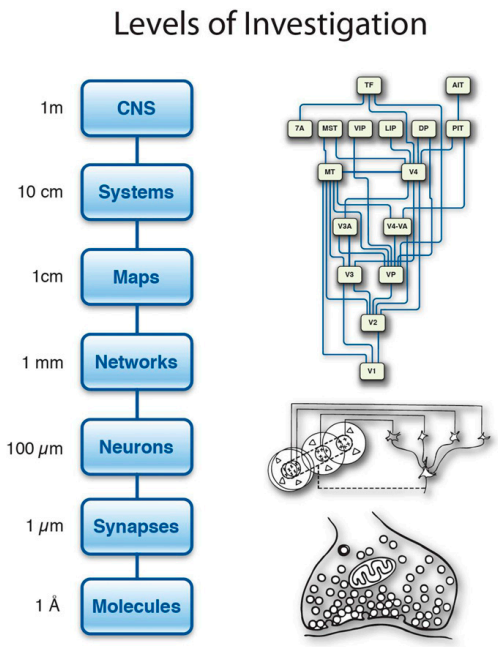

図5　脳を調べるレベル。シナプスでの少数の分子による信号伝達でエネルギー効率を上げる。脳内のニューロン間の相互接続は3次元的である。局所的には接続性が高いが、離れた皮質領域間では比較的疎である。大脳皮質の組織化原理は、感覚面と運動面の複数のマップが階層化されていることに基づいている。大脳皮質は、多くの皮質下領域と連携し、行動を生み出す中枢神経系 (CNS) を形成する。

脳には 11 桁の空間的に構造化された計算機部品がある(図5)。シナプスのレベルでは、米粒ほどの大きさの大脳皮質の 1 立方ミリメートルあたり、10 億個のシナプスが存在する。現在の最大の深層学習ネットワークは 10 億個の重みに達する。大脳皮質には問題解決に特化した何十万もの深層学習ネットワークに相当する力がある。これらのエキスパートネットワークはどのように構成されているのだろうか？ネットワークレベル以上の研究レベルでは大脳皮質の異なる領域間の情報の流れを整理している。だがこれはシステムレベルの通信問題である。大脳皮質内の情報の流れがどのように管理されているかを研究することで何千もの専門的なネットワークをどのように組織化するかについて多くのことを学ぶことができる。大脳皮質内の長距離接続は、情報を長距離に送るために必要なエネルギーや、大きなスペースを必要とすることから、コストがかかるため疎らになっている。スイッチングネットワークは、感覚野と運動野の間で情報を伝達し、進行中の認知的要求に応じて迅速に再構成することができる(17)。

また次世代の AI システムを構築する上で大きな課題となるのが、高度に異種混合された深層学習専門家ネットワークのシステムのメモリ管理である。すでに学習した記憶を劣化させることなく、柔軟にネットワークを更新する必要があり、これが安定した生涯学習の維持の問題である(20)。記憶の喪失やサブシステム間の干渉を最小限にする方法がある。一つの方法は、新しい経験を保存する場所を選択することである。これは、大脳皮質がグローバルにまとまった電気的活動パターンに入る睡眠中に起こる。睡眠紡錘体と呼ばれる短時間の振動現象は、夜間に何千回も繰り返され、記憶の定着に関連している。睡眠紡錘体は、日中に経験した最近のエピソードの再生によって引き起こされ、大脳皮質の長期意味記憶に解析的に統合される (21,22)。

4　深層学習の未来

今日の深層学習は大脳皮質に着想を得たものである。動作や生命活動を制御するためには、もっと幅広いアーキテクチャが必要である。強化学習を司る大脳基底核や、運動命令の前方モデルを脳に提供する小脳など、生存に不可欠な皮質下の部分は、すべての脊椎動物に見られる。人間は超社会性を備えており、複雑な社会的相互作用を支える大脳皮質および皮質下の神経回路が広範囲に存在する(23)。これらの脳領域は、自律的な AI システムの構築を目指す人々にインスピレーションを与えるだろう。

例えば脳幹のドーパミンニューロンは報酬予測誤差を計算する。だが、これは強化学習における時間差学習アルゴリズムの重要な計算であり, 深層学習との連携により, 2017 年に AlphaGo が囲碁の世界チャンピオンである柯潔を倒す原動力となった (24, 25)。大脳皮質と基底核全体に拡散して投射する中脳のドーパミンニューロンからの記録は、シナプス可塑性を調節し、長期的な報酬を得るための動機付けとなる (26)。その後、人間におけるドーパミンニューロンの役割が確認されたことで、人間がどのように経済的意思決定を行うのかをより深く理解することを目的とした神経経済学という新しい分野が生まれた(27)。他のいくつかの神経調節系も負の報酬, 驚き, 自信, 時間的割引など、行動を導くために脳の全体的な状態を制御している(28)。

運動系もまた生物学的にインスピレーションを得たソリューションが役立つ可能性のある AI の分野である。動物の動きは流動的で一般的なロボットの硬い動きとは異なる。この違いはすべての動物が高次元の筋肉組織を制御する際に見せる並外れた柔軟性にある。高次元の運動計画空間での協調行動は深層学習ネットワークで活発に研究されている分野である(29)。また脊髄, 脳幹, 前脳の複数の制御層がどのように協調するかを説明する分散制御理論も必要である。脳も制御システムも、フィードバックループの時間的な遅れに対処しなければならず、不安定になる可能性がある。小脳にある身体の前方モデルは、運動コマンドの感覚的な結果を予測する方法を提供し、感覚的な予測誤差はオープンループ制御を最適化するために使用される。例えば、前庭眼球反射 (VOR) は、頭

部の加速度信号を開ループで迅速に利用することで、頭部の動きにもかかわらず網膜上の画像を安定させる。 VOR のゲインは、網膜からのスリップ信号によって適応され、小脳はそのスリップを軽減するために利用します (30)。脳には感覚神経や運動神経の帯域幅が限られているという制約があるが、速度と精度のトレードオフが多様な構成要素を持つ層状制御システムであれば、これを克服できる(31)。同様の多様性は人工的なシステムにも存在し不完全なコンポーネントを持つにもかかわらず高速で正確な制御を可能にしている(32)。

5　人工一般知能へ向けて

深層学習の現状から人工的な一般知能への道はあるのか？進化の観点から見ると、ほとんどの動物はそれぞれのニッチで生き残るために必要な問題を解決することができる。だが, 一般的な抽象的推論は人間の系統ではより最近になって出現した。しかし人間はあまり得意ではなく論理的に推論できるようになるには長い訓練が必要となる。これは論理に最適化されていない脳のシステムを使って、論理的なステップをシミュレートしているからである。小学生は簡単な算数をマスターするために何年もかけて勉強する。これは 1 秒単位のクロックを持つデジタルコンピュータを模したものである。とはいえ人間の推論は合理的な計画や意思決定のために深層学習ネットワークの大規模システムを進化させることが可能であるはずだという原理的な証明でもある。しかし DeepMind 社が開発したコピー, ソート, ナビゲートの方法を学習するニューラル・チューリング・マシーンのように、ハイブリッドなソリューションも可能かもしれない(33)。オルゲルの第 2 法則によれば、自然は我々よりも賢いが改善はまだ可能かもしれない。

近年ディープネットワークにおける教師付き学習の成功により大規模なデータセットが利用できるアプリケーションが急増している。言語翻訳は翻訳されたテキストの大規模なコーパスで学習することで大きく改善された。しかし大規模なラベル付きデータセットが利用できないアプリケーションも多くある。人間は物理的な世界の結果を無意識のうちに予測し、予想外のことに驚くものである。他のデータストリームから将来の出力を予測することを学習の目的とする自己教師付き学習は、有望な方向性である (34)。また、模倣学習は、重要な行動を学び、世界についての知識を得るための強力な手段である (35)。人間には様々な学習方法があり成人レベルのパフォーマンスを得るためには、長い期間の発達が必要である。

脳は知的にそして自発的にアイデアや問題解決策を生み出す。脳スキャナーの中で静かに横になってもらうと活動は感覚運動領域から無意識の活動を含む内的思考をサポートする領域のデフォルトモードネットワークに切り替わる。生成ニューラルネットワークモデルは豊富にある生の感覚データから結合確率分布を学習することを目的としており、教師なしで学習することができる。ボルツマンマシンは生成モデルの一例である(8)。ボルツマンマシンは入力を分類するように学習された後、出力ユニットをオンにすると、そのカテゴリの例のシーケンスが入力層に生成される(36)。また生成的逆問題ネットワークは自己教師付き学習で学習した確率分布から新しいサンプルを生成することもできる(37)。また、脳は、夢を見ている間に、しばしば奇妙な視覚イメージを生成する。

6　今後の展開

我々は「情報の時代」と呼ばれる新しい時代の始まりを迎えている。データを情報に、情報を知識に、知識を理解に、そして運が良ければ知識を知恵に変えるパイプラインの源となるセンサーからデータが湧き出ている。我々は現実の世界で複雑な高次元問題を扱うための最初の一歩を踏み出した。赤ちゃんのように、歩幅よりもつまずきが多いが、重要なのは、私たちが正しい方向に向かっているということである。深層学習ネットワークは、デジタルコンピューターと現実世界の架け橋である。これにより、私たちは自分の言葉でコンピューターとコミュニケーションをとることができる。我々はすでにスマートスピーカーと会話している。スマートスピーカはさらに賢くなる。キーボードは廃止されタイプライタと並んで博物館に収蔵されることになるでだろう。これにより、ディープラーニングの恩恵を誰もが受けられるようになるだろう。

ユージン・ウィグナーは、「自然科学における数学の理不尽な有効性」という論文の中で物理理論の数学的構造からその理論に対する深い洞察が得られそれが経験的な予測につながることが多いことに驚嘆している(38)。また、物理定数と呼ばれる方程式のパラメータが非常に少ないことにも注目している。この記事のタイトルはウィグナーの論文と同じである。しかし物理法則とは異なり、深層学習ネットワークには豊富なパラメータが存在し、それらは可変である。私たちは、超高次元空間での表現や最適化を模索し始めたばかりである。いつの日か、深層学習ネットワークの構造を分析することで、理論的な予測が可能になり、知能の本質についての深い洞察が得られるかもしれない。私たちは次元の恩恵を受けることができるのである。

世界の信号の複雑さを記述する関数のクラスを 1 つ見つけたがおそらく他にもあるだろう。私たちが住んでいる 3 次元の世界や、デジタルコンピュータの 1 次元の命令列から得られる直感を超えた、高次元空間における超並列アルゴリズム

の世界が、私たちはまだ探求していないのかもしれない。フラットランドの正方形の紳士 (図1) やフラマリオンの彫刻の探検家 (図6) のように我々は古い地平線をはるかに超えた新しい世界を垣間見たのである。

1. E. A. Abbott, *Flatland: A Romance in Many Dimensions* (Seeley & Co., London, 1884).
2. L. Breiman, Statistical modeling: The two cultures. *Stat. Sci.* **16**, 199–231 (2001).
3. N. Chomsky, *Knowledge of Language: Its Nature, Origins, and Use* (Convergence, Praeger, Westport, CT, 1986).
4. T. J. Sejnowski, *The Deep Learning Revolution: Artificial Intelligence Meets Human Intelligence* (MIT Press, Cambridge, MA, 2018).
5. F. Rosenblatt, *Perceptrons and the Theory of Brain Mechanics* (Cornell Aeronautical Lab Inc., Buffalo, NY, 1961), vol. VG-1196-G, p. 621.
6. W. S. McCulloch, W. H. Pitts, A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.* **5**, 115–133 (1943).
7. M. Minsky, S. Papert, *Perceptrons* (MIT Press, Cambridge, MA, 1969).
8. D. H. Ackley, G. E. Hinton, T. J. Sejnowski, A learning algorithm for Boltzmann Machines. *Cogn. Sci.* **9**, 147–169 (1985).
9. T. J. Sejnowski, The book of Hebb. *Neuron* **24**, 773–776 (1999).
10. D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
11. R. Pascanu, Y. N. Dauphin, S. Ganguli, Y. Bengio, On the saddle point problem for non-convex optimization. arXiv:1405.4604 (19 May 2014).
12. P. L. Bartlett, P. M. Long, G. Lugosi, A. Tsigler, Benign overfitting in linear regression. arXiv:1906.11300 (26 June 2019).
13. T. Poggio, A. Banburski, Q. Liao, Theoretical issues in deep networks: Approximation, optimization and generalization. arXiv:1908.09375 (25 August 2019).
14. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, "Distributed representations of words and phrases and their compositionality" in *Proceedings of the 26th International Conference on Neural Imaging Processing Systems* (Curran Associates, 2013), vol. 2, pp. 3111–3119.
15. D. McCullough, *The Wright Brothers* (Simon & Schuster, New York, 2015).
16. S. Navlakha, Z. Bar-Joseph, Algorithms in nature: The convergence of systems biology and computational thinking. *Mol. Syst. Biol.* **7**, 546 (2011).
17. S. B. Laughlin, T. J. Sejnowski, Communication in neuronal networks. *Science* **301**, 1870–1874 (2003).
18. K. Zhang, T. J. Sejnowski, A universal scaling law between gray matter and white matter of cerebral cortex. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 5621–5626 (2000).
19. S. Srinivasan, C. F. Stevens, Scaling principles of distributed circuits. *Curr. Biol.* **29**, 2533–2540.e7 (2019).
20. G. Gary Anthes, Lifelong learning in artificial neural networks. *Commun. ACM* **62**, 13–15 (2019).
21. L. Muller *et al*., Rotating waves during human sleep spindles organize global patterns of activity during the night. *eLife* **5**, 17267 (2016).
22. R. Todorova, M. Zugaro, Isolated cortical computations during delta waves support memory consolidation. *Science* **366**, 377–381 (2019).
23. P. S. Churchland, *Conscience: The Origins of Moral Intuition* (W. W. Norton, New York, 2019).
24. Wikipedia, AlphaGo versus Ke Jie. https://en.wikipedia.org/wiki/AlphaGo_versus_Ke_Jie. Accessed 8 January 2020.
25. D. Silver *et al*., A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science* **362**, 1140–1144 (2018).
26. P. R. Montague, P. Dayan, T. J. Sejnowski, A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J. Neurosci.* **16**, 1936–1947 (1996).
27. P. W. Glimcher, C. Camerer, R. A. Poldrack, E. Fehr, *Neuroeconomics: Decision Making and the Brain* (Academic Press, New York, 2008).
28. E. Marder, Neuromodulation of neuronal circuits: Back to the future. *Neuron* **76**, 1–11 (2012).
29. I. Akkaya *et al*., Solving Rubik's cube with a robot hand. arXiv:1910.07113 (16 October 2019).
30. S. du Lac, J. L. Raymond, T. J. Sejnowski, S. G. Lisberger, Learning and memory in the vestibulo-ocular reflex. *Annu. Rev. Neurosci.* **18**, 409–441 (1995).
31. Y. Nakahira, Q. Liu, T. J. Sejnowski, J. C. Doyle, Fitts' Law for speed-accuracy trade-off describes a diversity-enabled sweet spot in sensorimotor control. arXiv:1906.00905 (18 September 2019).
32. Y. Nakahira, Q. Liu, T. J. Sejnowski, J. C. Doyle, Diversity-enabled sweet spots in layered architectures and speed-accuracy trade-offs in sensorimotor control. arXiv: 1909.08601 (18 September 2019).
33. A. Graves, G. Wayne, I. Danihelka, Neural turing machines. arXiv:1410.540 (20 October 2014).
34. A. Rouditchenko, H. Zhao, C. Gan, J. McDermott, A. Torralba, Self-supervised audio-visual co-segmentation. arXiv:1904.09013 (18 April 2019).
35. S. Schaal, Is imitation learning the route to humanoid robots? *Trends Cogn. Sci.* **3**, 233–242 (1999).
36. G. E. Hinton, S. Osindero, Y. Teh, A fast learning algorithm for deep belief nets. *Neural Comput.* **18**, 1527–1554 (2006).
37. I. J. Goodfellow *et al*., Generative adversarial nets. arXiv:1406.2661(10 June 2014).
38. E. P. Wigner, The unreasonable effectiveness of mathematics in the natural sciences. Richard Courant lecture in mathematical sciences delivered at New York University, May 11, 1959. *Commun. Pure Appl. Math.* **13**, 1–14 (1960).