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 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.

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.

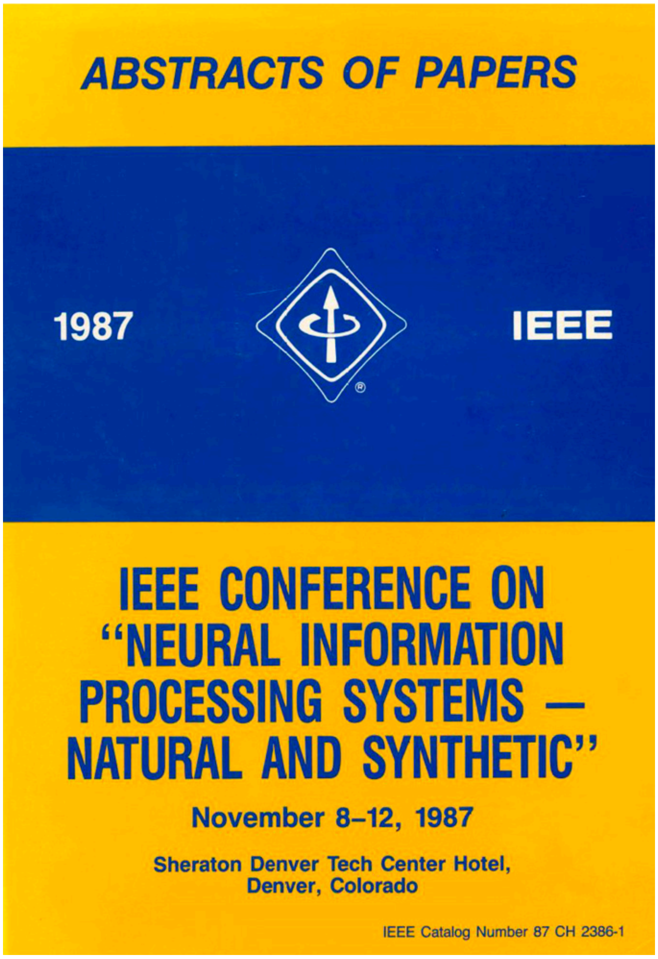
enough examples, his learning algorithm was guaranteed to find it. The learning algorithm used labeled data to make small changes to parameters, which were the weights on the inputs to a binary threshold unit, implementing gradient descent. This simple paradigm is at the core of much larger and more 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,



ABSTRACTS OF PAPERS

1987 IEEE

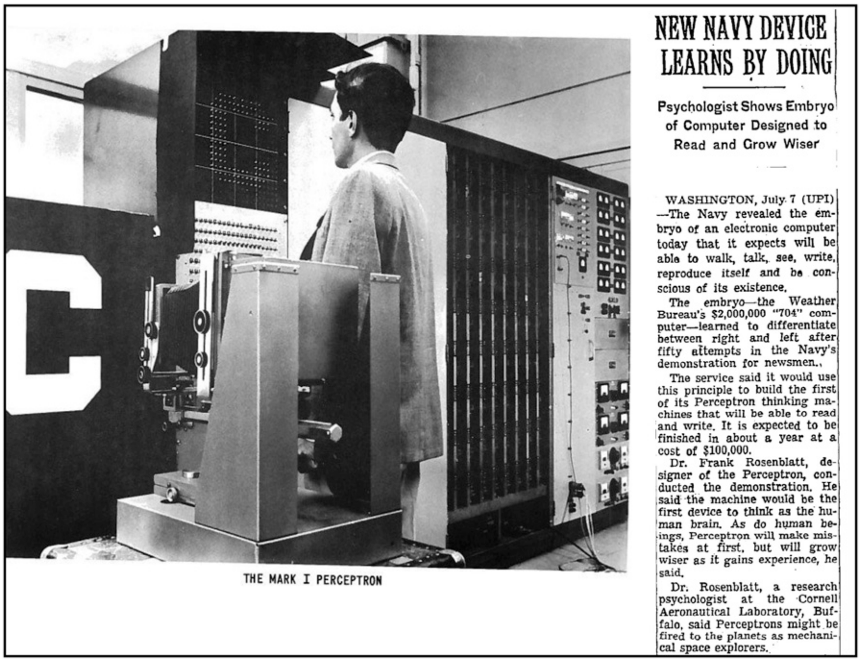
IEEE CONFERENCE ON
"NEURAL INFORMATION
PROCESSING SYSTEMS —
NATURAL AND SYNTHETIC"

November 8–12, 1987

Sheraton Denver Tech Center Hotel,
Denver, Colorado

IEEE Catalog Number 87 CH 2386-1

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.



NEW NAVY DEVICE LEARNS BY DOING

Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser

WASHINGTON, July 7 (UPI)—The Navy revealed the embryo of an electronic computer today that it expects will be able to walk, talk, see, write, reproduce itself and be conscious of its existence.

The embryo—the Weather Bureau's $2,000,000 "704" computer—learned to differentiate between right and left after fifty attempts in the Navy's demonstration for newsmen.,

The service said it would use this principle to build the first of its Perceptron thinking machines that will be able to read and write. It is expected to be finished in about a year at a cost of $100,000.

Dr. Frank Rosenblatt, designer of the Perceptron, conducted the demonstration. He said the machine would be the first device to think as the human brain. As do human beings, Perceptron will make mistakes at first, but will grow wiser as it gains experience, he said.

Dr. Rosenblatt, a research psychologist at the Cornell Aeronautical Laboratory, Buffalo, said Perceptrons might be fired to the planets as mechanical space explorers.

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.

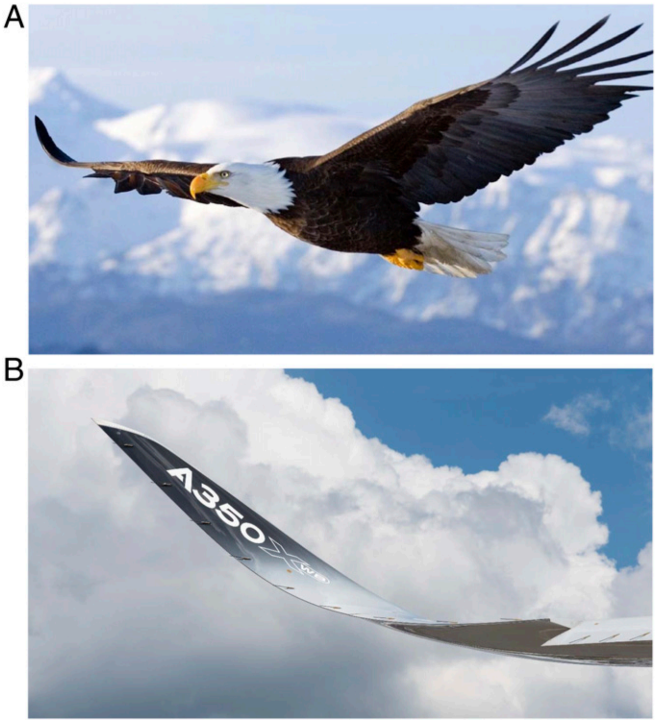


Fig. 4. Nature has optimized birds for energy efficiency. (*A*) The curved feathers at the wingtips of an eagle boosts energy efficiency during gliding. (*B*) Winglets on a commercial jets save fuel by reducing drag from vortices.

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

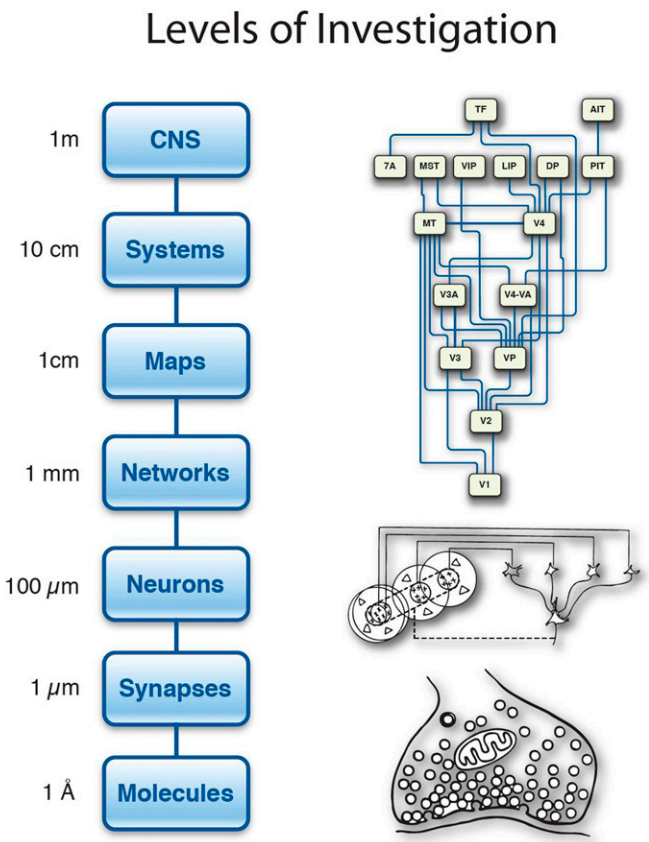


Fig. 5. Levels of investigation of brains. Energy efficiency is achieved by signaling with small numbers of molecules at synapses. Interconnects between neurons in the brain are 3D. Connectivity is high locally but relatively sparse between distant cortical areas. The organizing principle in the cortex 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

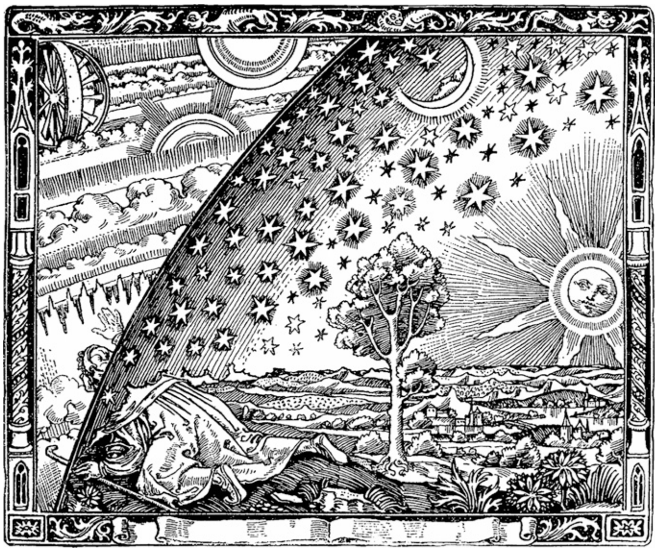


Fig. 6. The caption that accompanies the engraving in Flammarion's book reads: "A missionary of the Middle Ages tells that he had found the point where the sky and the Earth touch …." Image courtesy of Wikimedia Commons/Camille Flammarion.

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).