Copyrighted material

Contents

Part One
What Is Learning?

Part Two
How Our Brain Learns

Part Three
The Four Pillars of Learning

Copyrighted material

About the Author

Stanislas Dehaene is one of Europe’s leading neuroscientists, and has been studying how education changes our brains for over thirty years. He is professor of Experimental Cognitive Psychology at the Collège de France, and director of the NeuroSpin brain imaging in Saclay. He is a member of seven academies and has received several international prizes, including the highest award in neuroscience, the Brain Prize. Dehaene’s previous books, which have been translated into fifteen languages, include *Consciousness and the Brain*, *Reading in the Brain* and *The Number Sense*.

Copyrighted material

PENGUIN BOOKS

HOW WE LEARN

‘His reach across the sciences is immense ... From Descartes to machine learning, Dehaene draws repeated parallels between computer science and the near-infinite complexity of the human brain’ Alex Quigly, *TES*

‘An expert overview of learning ... Never mind our opposable thumb, upright posture, fire, tools, or language; it is education that enabled humans to conquer the world ... Dehaene’s fourth insightful exploration of neuroscience will pay dividends’ *Kirkus Reviews*

‘Dehaene rigorously examines our remarkable capacity for learning. The baby brain is especially awesome and not a “blank slate” ... Dehaene’s portrait of the human brain is fascinating’ *Booklist*

‘Richly instructive for educators, parents and others interested in how to most effectively foster the pursuit of knowledge’ *Publishers Weekly*

‘Dehaene’s tall task to present contributions of brain science to the way we practice education is the gem of his book ... a potent antidote against the threat of neuromyths ... the best presentation card that the field of educational neuroscience currently has’ Emmanuel Rosado, LWOS Life

Copyrighted material

For Aurore, who was born this year,

and for all those who once were babies.

Copyrighted material

Begin by making a more careful study of your pupils, for it is clear that you know nothing about them.

Jean-Jacques Rousseau, *Emile, or On Education* (1762)

This is a strange and amazing fact: we know every nook and cranny of the human body, we have catalogued every animal on the planet, we have described and baptized every blade of grass, but we have left psychological techniques to their empiricism for centuries, as if they were of lesser importance than those of the healer, the breeder or the farmer.

Jean Piaget, “La pédagogie moderne” (1949)

If we don’t know how we learn, how on earth do we know how to teach?

L. Rafael Reif, president of MIT (March 23, 2017)

Copyrighted material

Credits

INSERT CREDITS

Insert Figure 1: copyright © by Nicolás Sainz Trápaga. Reproduced with permission.

Insert Figure 2, top: from the Google Brain Team. “Using Machine Learning to Explore Neural Network Architecture.” *Google AI Blog* (2017). https://ai.googleblog.com/2017/05/using-machine-learning-to-explore.html.

Insert Figure 2, bottom: from Olah, Chris, Alexander Mordvintsev, and Ludwig Schubert. “Feature Visualization.” *Distill* (2017). https://distill.pub/2017/feature-visualization/. Licensed under Creative Commons Attribution License CC-BY 4.0.

Insert Figure 3, right: from Guerguiev, Jordan, Timothy P. Lillicrap, and Blake A. Richards. “Towards deep learning with segregated dendrites.” *ELife*, 6, e22901, (2017). https://elifesciences.org/articles/22901. Licensed under Creative Commons Attribution License CC-BY 4.0.

Insert Figure 3, left: from the MNIST database of handwritten digits. LeCun, Yann, Corinna Cortes, and Christopher J. C. Burges. http://yann.lecun.com/exdb/mnist/.

Insert Figure 4: from figures 2 and 3 in Kemp, Charles, and Joshua B. Tenenbaum. “The discovery of structural form.” *Proceedings of the National Academy of Sciences of the United States of America*, 105(31), 10687–10692 (2008). https://www.pnas.org/content/105/31/10687.short. Copyright © 2008 by National Academy of Sciences, U.S.A.

Insert Figure 5, top left: courtesy of Fei Xu Lab.

Insert Figure 5, top right: courtesy of Moira Dillon and Elizabeth Spelke.

Insert Figure 6, top: courtesy of G. Dehaene-Lambertz and J. Dubois.

Insert Figure 6, bottom: redrawn from data in Dehaene-Lambertz, Ghislaine, Lucie Hertz-Pannier, Jessica Dubois, Sébastien Mériaux, Alexis Roche, Mariano Sigman, and Stanislas Dehaene. “Functional organization of perisylvian activation during presentation of sentences in preverbal infants.” *Proceedings of the National Academy of Sciences of the United States of America*, 103(38), 14240–14245, (2006). https://www.pnas.org/content/103/38/14240. Copyright © 2006 by National Academy of Sciences, U.S.A.

Insert Figure 7: modified with permission from a figure provided courtesy of Leah Krubitzer. For review of the corresponding research, see Krubitzer, Leah. “The Magnificent Compromise: Cortical Field Evolution in Mammals.” *Neuron*, 56(2), 201–208, (2007).

Insert Figure 8, top: modified with permission from a figure provided by Alain Chédotal. See Belle, Morgane, David Godefroy, Gérard Couly, Samuel A. Malone, Francis Collier, Paolo Giacobini, and Alain Chédotal. “Tridimensional Visualization and Analysis of Early Human Development.” *Cell*, vol. 169(1), 161-173.e12, (2017). https://doi.org/10.1016/j.cell.2017.03.008.

Insert Figure 8, bottom: courtesy of G. Dehaene-Lambertz and J. Dubois.

Copyrighted material

maintenance of a lost first language." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111(48), pg. 17314–17319, (2014). https://www.pnas.org/content/111/48/17314.

Figure Page 114, top: photo courtesy of Eric Knudsen.

Figure Page 114, bottom: from figures 2 and 3 in Knudsen, Eric I., Weimin Zheng, and William M. DeBello. "Traces of learning in the auditory localization pathway." *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97(22), pg.11815–11820, (2000). https://www.pnas.org/content/97/22/11815. Copyright © 2000 by the National Academy of Sciences, U.S.A.

Figure Page 117, top: copyright © 2001 by Michael Carroll. Reproduced with permission.

Figure Page 117, bottom: adapted from figure 1 in Almas, Alisa N., Kathryn A. Degnan, Anca Radulescu, Charles A. Nelson III, Charles H. Zeanah, and Nathan A. Fox. "Effects of early intervention and the moderating effects of brain activity on institutionalized children's social skills at age 8." *Proceedings of the National Academy of Sciences of the United States of America,* vol. 109 Suppl 2, pg. 17228–17231, (2012). https://www.pnas.org/content/109/Supplement_2/17228.

Figure Page 135: figure created by the author, from data published in Dehaene, Stanislas, Felipe Pegado, Lucia W. Braga, Paulo Ventura, Gilberto Nunes Filho, Antoinette Jobert, Ghislaine Dehaene-Lambertz, Régine Kolinsky, José Morais, and Laurent Cohen. "How Learning to Read Changes the Cortical Networks for Vision and Language." *Science*, vol. 330(6009), pg. 1359–1364, (2010). https://doi.org/10.1126/science.1194140.

Figure Page 137: figure adapted from Dehaene-Lambertz, Ghislaine, Karla Monzalvo, and Stanislas Dehaene (2018). "The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition." *PLoS Biology*, vol. 16(3), e2004103, (2018). https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.2004103. Licensed under Creative Commons Attribution License CC-BY 4.0.

Figure Page 149: from Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention." ArXiv:1502.03044 [Cs], (2015). Retrieved from http://arxiv.org/abs/1502.03044.

Figure Page 157: figure composed by the author, based on graphs provided courtesy of Bruce McCandliss, from data reported in Yoncheva, Y. N., Blau, V. C., Maurer, U., & McCandliss, B. D. "Attentional Focus During Learning Impacts N170 ERP Responses to an Artificial Script." *Developmental Neuropsychology*, 35(4), 423–445 (2010). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4365954/.

Figure Page 166, top: copyright © by Stanislas Dehaene.

Figure Page 166, bottom: adapted with permission of Robert Zatorre, from data in Bermudez, Patrick, Jason P. Lerch, Alan C. Evans, and Robert J. Zatorre. "Neuroanatomical Correlates of Musicianship as Revealed by Cortical Thickness and Voxel-Based Morphometry." *Cereb Cortex*, vol. 19(7), pg. 1583–1596, (2009). https://academic.oup.com/cercor/article/19/7/1583/317010.

Figure Page 170, top: composed by the authors, based on photographs provided courtesy of György Gergely. Data from Egyed, Katalin, Ildikó Király, and György Gergely. "Communicating Shared Knowledge in Infancy." *Psychological Science*, vol. 24(7), pg. 1348–1353, (2013). https://journals.sagepub.com/doi/10.1177/0956797612471952.

Copyrighted material

Figure Page 170, bottom: composed with data from Gergely, György, Harold Bekkering, and Ildikó Király. “Rational imitation in preverbal infants.” *Nature*, vol. 415(6873), pg. 755 (2002). https://www.nature.com/articles/415755a.

Figure Page 192: adapted from figure 3 in Kaplan, Frederic, and Pierre-Yres Oudeyer. “In Search of the Neural Circuits of Intrinsic Motivation.” *Frontiers in Neuroscience*, 1(1), 225, (2007). https://www.frontiersin.org/articles/10.3389/neuro.01.1.1.017.2007/full. Copyright © 2007 by Kaplan and Oudeyer. This is an open-access article subject to an exclusive license agreement between the authors and the Frontiers Research Foundation, which permits unrestricted use, distribution, and reproduction in any medium, provided the original authors and source are credited. Licensed under Creative Commons Attribution License CC-BY 4.0.

Figure Page 217: copyright © by Stanislas Dehaene.

Attribution 4.0 International (CC BY 4.0) can be found at https://creativecommons.org/licenses/by/4.0/.

Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) can be found at https://creativecommons.org/licenses/by-sa/4.0/.

Copyrighted material

Introduction

IN SEPTEMBER 2009, AN EXTRAORDINARY CHILD FORCED ME TO DRASTICALLY revise my ideas about learning. I was visiting the Sarah Hospital in Brasilia, a neurological rehabilitation center with a white architecture inspired by Oscar Niemeyer, with which my laboratory has collaborated for about ten years. The director, Lucia Braga, asked me to meet one of her patients, Felipe, a young boy only seven years old, who had spent more than half his life in a hospital bed. She explained to me how, at the age of four, he had been shot in the street—unfortunately not such a rare event in Brazil. The stray bullet had severed his spinal cord, thus rendering him almost completely paralyzed (tetraparetic). It also destroyed the visual areas of his brain: he was fully blind. To help him breathe, an opening was made in his trachea, at the base of his neck. And for over three years, he had been living in a hospital room, locked within the coffin of his inert body.

In the corridor leading to his room, I remember bracing myself at the thought of having to face a broken child. And then I meet ... Felipe, a lovely little boy like any other seven-year-old—talkative, full of life, and curious about everything. He speaks flawlessly with an extensive vocabulary and asks me mischievous questions about French words. I learn that he has always been passionate about languages and never misses an opportunity to enrich his trilingual vocabulary (he speaks Portuguese, English, and Spanish). Although he is blind and bedridden, he escapes into his imagination by writing his own novels, and the hospital team has encouraged him in this path. In a few months, he learned to dictate his stories to an assistant, then write them himself using a special keyboard connected to a computer and sound card. The pediatricians and speech therapists take turns at his bedside, transforming his writings into real, tactile books with embossed illustrations that he proudly sweeps with his fingers, using the little sense of touch that he has left. His stories speak of heroes and heroines, mountains and lakes that he will never see, but that he dreams of like any other little boy.

Meeting with Felipe deeply moved me, and also persuaded me to take a closer look at what is probably the greatest talent of our brain: the ability to learn. Here was a child whose very existence poses a challenge to neuroscience. How do our brain's cognitive faculties resist such a radical upheaval of their environment? Why could Felipe and I share the same thoughts, given our extraordinarily different sensory experiences? How do different human brains converge on the same concepts, almost regardless of how and when they learn them?

Many neuroscientists are empiricists: together, with the English Enlightenment philosopher John Locke (1632–1704), they presume that the brain simply draws its knowledge from its environment. In this view, the main property of cortical circuits is their plasticity, their ability to adapt to their inputs. And, indeed, nerve cells possess a remarkable ability to constantly adjust their synapses according to the signals they receive. Yet if this were the brain's main drive, my little Felipe, deprived of visual and

Copyrighted material

motor inputs, should have become a profoundly limited person. By what miracle did he manage to develop strictly normal cognitive abilities?

Felipe's case is by no means unique. Everybody knows the story of Helen Keller (1880–1968) and Marie Heurtin (1885–1921), both of whom were born deaf and blind and yet, after years of grueling social isolation, learned sign language and ultimately became brilliant thinkers and writers.[1] Throughout these pages, we will meet many other individuals who, I hope, will radically alter your views on learning. One of them is Emmanuel Giroux, who has been blind since the age of eleven but became a top-notch mathematician. Paraphrasing the fox in Antoine de Saint-Exupéry's *The Little Prince* (1943), Giroux confidently states: "In geometry, what is essential is invisible to the eye. It is only with the mind that you can see well." How does this blind man manage to swiftly navigate within the abstract spaces of algebraic geometry, manipulating planes, spheres, and volumes without ever seeing them? We will discover that he uses the same brain circuits as other mathematicians, but that his visual cortex, far from remaining inactive, has actually repurposed itself to do math.

I will also introduce you to Nico, a young painter who, while visiting the Marmottan Museum in Paris, managed to make an excellent copy of Monet's famous painting *Impression, Sunrise* (see figure 1 in the color insert). What is so exceptional about this? Nothing, besides the fact that he accomplished it with only a single hemisphere, his left one—the right half of his brain was almost fully removed at the age of three! Nico's brain learned to squeeze all his talents into half a brain: speech, writing, and reading, as usual, but drawing and painting too, which are generally thought to be functions of the right hemisphere, and also computer science and even wheelchair fencing, a sport in which he has reached the rank of champion in Spain. Forget everything you were told about the respective roles of both hemispheres, because Nico's life proves that anyone can become a creative and talented artist without a right hemisphere! Cerebral plasticity seems to work miracles.

We will also visit the infamous orphanages of Bucharest where children were left from birth in quasi-abandon—and yet, years later, some of them, adopted before the age of one or two, have had almost normal school experiences.

All these examples illustrate the extraordinary resilience of the human brain: even major trauma, such as blindness, the loss of a hemisphere, or social isolation, cannot extinguish the spark of learning. Language, reading, mathematics, artistic creation: all these unique talents of the human species, which no other primate possesses, can resist massive injuries, such as the removal of a hemisphere or the loss of sight and motor skills. Learning is a vital principle, and the human brain has an enormous capacity for plasticity—to change itself, to adapt. Yet we will also discover dramatic counterexamples, where learning seems to freeze and remain powerless. Consider pure alexia, the inability to read a single word. I have personally studied several adults, all of whom were excellent readers, who had a tiny stroke restricted to a minuscule brain area that rendered them incapable of deciphering words as simple as "dog" or "mat." I remember a brilliant trilingual woman, a faithful reader of the French newspaper *Le Monde*, who was deeply sorrowed at the fact that, after her brain injury, every page of the daily press looked like Hebrew. Her determination to relearn to read was at least as strong as the stroke that she had suffered was severe. However, after two years of perseverance, her reading level still did not exceed that of a kindergartner: it took her several seconds to read a single word, letter by letter, and she still stumbled on every word. Why couldn't she learn? And why do some children, who suffer from dyslexia,

Copyrighted material

dyscalculia, or dyspraxia, show a similar radical hopelessness in acquiring reading, calculating, or writing while others surf smoothly through those fields?

Brain plasticity almost seems temperamental: sometimes it overcomes massive difficulties, and other times it leaves children and adults who are otherwise highly motivated and intelligent with debilitating disabilities. Does it depend on particular circuits? Do these circuits lose their plasticity over the years? Can plasticity be reopened? What are the rules that govern it? How can the brain be so effective from birth and throughout a child's youth? What algorithms allow our brain circuits to form a representation of the world? Would understanding them help us learn better and faster? Could we draw inspiration from them in order to build more efficient machines, artificial intelligences that would ultimately imitate us or even surpass us? These are some of the questions that this book attempts to answer, in a radically multidisciplinary manner, drawing on recent scientific discoveries in cognitive science and neuroscience, but also in artificial intelligence and education.

WHY LEARN?

Why do we have to learn in the first place? The very existence of the capacity to learn raises questions. Wouldn't it be better for our children to immediately know how to speak and think, right from day one, like Athena, who, according to legend, emerged into the world from Zeus's skull, fully grown and armed, as she let out her war cry? Why aren't we born pre-wired, with pre-programmed software and exactly the pre-loaded knowledge necessary to our survival? In the Darwinian struggle for life, shouldn't an animal who is born mature, with more knowledge than others, end up winning and spreading its genes? Why did evolution invent learning in the first place?

My answer is simple: a complete pre-wiring of the brain is neither possible nor desirable. Impossible, really? Yes, because if our DNA had to specify all the details of our knowledge, it simply would not have the necessary storage capacity. Our twenty-three chromosomes contain three billion pairs of the "letters" A, C, G, T—the molecules adenine, cytosine, guanine, and thymine. How much information does that represent? Information is measured in bits: a binary decision, 0 or 1. Since each of the four letters of the genome codes for two bits (we can code them as 00, 01, 10, and 11), our DNA therefore contains a total of six billion bits. Remember, however, that in today's computers, we count in bytes, which are sequences of eight bits. The human genome can thus be reduced to about 750 megabytes—the contents of an old-fashioned CD-ROM or a small USB key! And this basic calculation does not even take into account the many redundancies that abound in our DNA.

From this modest amount of information, inherited from millions of years of evolution, our genome, initially confined to a single fertilized egg, manages to set up the whole body plan—every molecule of every cell in our liver, kidneys, muscles, and, of course, our brain: eighty-six billion neurons, a thousand trillion connections .... How could our genome possibly specify each one of them? Assuming that each of our nerve connections encodes only one bit, which is certainly an underestimate, the capacity of our brain is on the order of one hundred terabytes (about 10^{15} bits), or a hundred thousand times more than the information in our genome. We are faced with a paradox: the fantastic palace that is our brain contains a hundred thousand times more detail than the architect's blueprints that are used to build it! I see only one explanation: the structural frame of the palace is built following the architect's guidelines (our genome), while the details are left to the project manager, who can adapt the blueprints to the terrain (the environment). Pre-wiring a human brain in all

Copyrighted material

Education magnifies the already considerable faculties of our brain—but could it perform even better? At school and at work, we constantly tinker with our brain's learning algorithms, yet we do so intuitively, without paying attention to how to learn. No one has ever explained to us the rules by which our brain memorizes and understands or, on the contrary, forgets and makes mistakes. It truly is a pity, because the scientific knowledge is extensive. An excellent website, put together by the British Education Endowment Foundation (EEF),[3] lists the most successful educational interventions – and it gives a very high ranking to the teaching of metacognition (knowing the powers and limits of one's own brain). Learning to learn is arguably the most important factor for academic success.

Fortunately, we now know a lot about how learning works. Thirty years of research, at the boundaries of computer science, neurobiology, and cognitive psychology, have largely elucidated the algorithms that our brain uses, the circuits involved, the factors that modulate their efficacy, and the reasons why they are uniquely efficient in humans. In this book, I will discuss all those points in turn. When you close this book, I hope you will know much more about your own learning processes. It seems fundamental, to me, that every child and every adult realize the full potential of his or her own brain and also, of course, its limits. Contemporary cognitive science, through the systematic dissection of our mental algorithms and brain mechanisms, gives new meaning to the famous Socratic adage "Know thyself." Today, the point is no longer just to sharpen our introspection, but to understand the subtle neuronal mechanics that generate our thoughts, in an attempt to use them in optimal accordance with our needs, goals, and desires.

The emerging science of how we learn is, of course, of special relevance to all those for whom learning is a professional activity: teachers and educators. I am deeply convinced that one cannot properly teach without possessing, implicitly or explicitly, a mental model of what is going on in the minds of the learners. What sort of intuitions do they start with? What steps do they have to take in order to move forward? What factors can help them develop their skills?

While cognitive neuroscience does not have all the answers, we begin to understand that all children start off life with a similar brain architecture—a *Homo sapiens* brain, radically different from that of other apes. I am not denying, of course, that our brains vary: the quirks of our genomes, as well as the whimsies of early brain development, grant us slightly different strengths and learning speeds. However, the basic circuitry is the same in all of us, as is the organization of our learning algorithms. There are therefore fundamental principles that any teacher must respect in order to be most effective. In this book, we will see many examples. All young children share abstract intuitions in the domains of language, arithmetic, logic, and probability, thus providing a foundation on which higher education must be grounded. And all learners benefit from focused attention, active engagement, error feedback, and a cycle of daily rehearsal and nightly consolidation—I call these factors the "four pillars" of learning, because, as we shall see, they lie at the foundation of the universal human learning algorithm present in all our brains, children and adults alike.

At the same time, our brains do exhibit individual variations, and in some extreme cases, a pathology can appear. The reality of developmental pathologies, such as dyslexia, dyscalculia, dyspraxia, and attention disorders, is no longer a subject of doubt. Fortunately, as we increasingly understand the common architecture from which these quirks arise, we also discover that simple strategies exist to detect and compensate for them. One of the goals of this book is to spread this growing scientific

Copyrighted material

knowledge, so that every teacher, and also every parent, can adopt an optimal teaching strategy. While children vary dramatically in *what* they know, they still share the same learning algorithms. Thus, the pedagogical tricks that work best with all children are also those that tend to be the most efficient for children with learning disabilities—they must be applied only with greater focus, patience, systematicity, and tolerance to error.

And the latter point is crucial: while error feedback is essential, many children lose confidence and curiosity because their errors are punished rather than corrected. In schools worldwide, error feedback is often synonymous with punishment and stigmatization—and later in this book I will have much to say about the role of school grades in perpetuating this confusion. Negative emotions crush our brain's learning potential, whereas providing the brain with a fear-free environment may reopen the gates of neuronal plasticity. There will be no progress in education without simultaneously considering the emotional and cognitive facets of our brain—in today's cognitive neuroscience, both are considered key ingredients of the learning cocktail.

THE CHALLENGE OF MACHINES

Today, human intelligence faces a new challenge: we are no longer the only champions of learning. In all fields of knowledge, learning algorithms are challenging our species' unique status. Thanks to them, smartphones can now recognize faces and voices, transcribe speech, translate foreign languages, control machines, and even play chess or Go—much better than we can. Machine learning has become a billion-dollar industry that is increasingly inspired by our brains. How do these artificial algorithms work? Can their principles help us understand what learning is? Are they already able to imitate our brains, or do they still have a long way to go?

While the current advances in computer science are fascinating, their limits are evident. Conventional deep learning algorithms mimic only a small part of our brain's functioning, the one that, I argue, corresponds to the first stages of sensory processing, the first two or three hundred milliseconds during which our brain operates in an unconscious manner. This type of processing is in no way superficial: in a fraction of a second, our brain can recognize a face or a word, put it in context, understand it, and even integrate it into a small sentence .... The limitation, however, is that the process remains strictly bottom-up, without any real capacity for reflection. Only in the subsequent stages, which are much slower, more conscious, and more reflective, does our brain manage to deploy all its abilities of reasoning, inference, and flexibility—features that today's machines are still far from matching. Even the most advanced computer architectures fall short of any human infant's ability to build abstract models of the world.

Even within their fields of expertise—for example, the rapid recognition of shapes—modern-day algorithms encounter a second problem: they are much less effective than our brain. The state of the art in machine learning involves running millions, even billions, of training attempts on computers. Indeed, machine learning has become virtually synonymous with big data: without massive data sets, algorithms have a hard time extracting abstract knowledge that generalizes to new situations. In other words, they do not make the best use of data.

In this contest, the infant brain wins hands down: babies do not need more than one or two repetitions to learn a new word. Their brain makes the most of extremely scarce data, a competence that still eludes today's computers. Neuronal learning algorithms often come close to optimal computation: they manage to extract the true

Copyrighted material

essence from the slightest observation. If computer scientists hope to achieve the same performance in machines, they will have to draw inspiration from the many learning tricks that evolution integrated into our brain: attention, for example, which allows us to select and amplify relevant information; or sleep, an algorithm by which our brain synthesizes what it learned on previous days. New machines with these properties are beginning to emerge, and their performance is constantly improving—they will undoubtedly compete with our brains in the near future.

According to an emerging theory, the reason that our brain is still superior to machines is that it acts as a statistician. By constantly attending to probabilities and uncertainties, it optimizes its ability to learn. During its evolution, our brain seems to have acquired sophisticated algorithms that constantly keep track of the uncertainty associated with what it has learned—and such a systematic attention to probabilities is, in a precise mathematical sense, the optimal way to make the most of each piece of information.[4]

Recent experimental data support this hypothesis. Even babies understand probabilities: from birth, they seem to be deeply embedded in their brain circuits. Children act like little budding scientists: their brains teem with hypotheses, which resemble scientific theories that their experiences put to the test. Reasoning with probabilities, in a largely unconscious manner, is deeply inscribed in the logic of our learning. It allows any of us to gradually reject false hypotheses and retain only the theories that make sense of the data. And, unlike other animal species, humans seem to use this sense of probabilities to acquire scientific theories from the outside world. Only *Homo sapiens* manages to systematically generate abstract symbolic thoughts and to update their plausibility in the face of new observations.

Innovative computer algorithms are beginning to incorporate this new vision of learning. They are called “Bayesian,” after the Reverend Thomas Bayes (1702–61), who outlined the rudiments of this theory as early as the eighteenth century. My hunch is that Bayesian algorithms will revolutionize machine learning—indeed, we will see that they are already able to extract abstract information with an efficiency close to that of a human scientist.

Our journey into the contemporary science of learning is a three-part trip.

In the first part, entitled “What Is Learning?”, we start by defining what it means for humans or animals—or indeed any algorithm or machine—to learn something. The idea is simple: to learn is to progressively form, in silicon and neural circuits alike, an internal model of the outside world. When I walk around a new town, I form a mental map of its layout—a miniature model of its streets and passageways. Likewise, a child who is learning to ride a bike is shaping, in her neural circuits, an unconscious simulation of how the actions on the pedals and handlebars affect the bike’s stability. Similarly, a computer algorithm learning to recognize faces is acquiring template models of the various possible shapes of eyes, noses, mouths, and their combinations.

But how do we set up the proper mental model? As we shall see, the learner’s mind can be likened to a giant machine with millions of tunable parameters whose settings collectively define what is learned (for instance, where the streets are likely to be in our mental map of the neighborhood). In the brain, the parameters are synapses, the connections between neurons, which can vary in strength; in most present-day computers, they are the tunable weights or probabilities that specify the strength of each tenable hypothesis. Learning, in both brains and machines, thus requires searching for an optimal combination of parameters that, together, define the mental model in every detail. In this sense, learning is a massive search problem—and in order

Copyrighted material

to understand how learning works in the human brain, it greatly helps to examine how learning algorithms operate in present-day computers.

By comparing the performance of computer algorithms with those of the brain, *in silico* versus *in vivo*, we will progressively get a sharper picture of what learning means at the brain level. For sure, mathematicians and computer scientists haven't managed to design learning algorithms as powerful as the human brain—yet. But they are beginning to home in on a theory of the optimal learning algorithm that any system should use if it aims for the greatest efficiency. According to this theory, the best learner operates as a scientist who makes rational use of probabilities and statistics. A new model emerges: that of the brain as a statistician, of cerebral circuits as computing with probabilities. This theory specifies a clear division of labor between nature and nurture: the genes first set up vast spaces of a priori hypotheses—and the environment then *selects* the hypotheses which best match the external world. The set of hypotheses is genetically specified; their selection is experience-dependent.

Does this theory correspond to how the brain works? And how is learning implemented in our biological circuits? What changes in our brains when we acquire a novel competence? In the second section, "How Our Brain Learns," we will turn to psychology and neuroscience. I will focus on babies, who are genuine learning machines without rivals. Recent data show that infants are indeed the budding statisticians predicted by the theory. Their remarkable intuition in the fields of language, geometry, numbers, and statistics confirms that they are anything but a blank slate, a tabula rasa. From birth, children's brain circuits are already organized and project hypotheses onto the outside world. But they also have a considerable margin of plasticity, which is reflected in the brain's perpetual effervescence of synaptic changes. Within this statistical machine, nature and nurture, far from opposing each other, join forces. The result is a structured yet plastic system with an unmatched ability to repair itself in the face of brain injury and to recycle its brain circuits in order to acquire skills unanticipated by evolution, such as reading or mathematics.

In the third part, "The Four Pillars of Learning," I detail some of the tricks that make our brain the most effective learning device known today. Four essential mechanisms, or "pillars," massively modulate our ability to learn. The first is attention: a set of neural circuits that select, amplify, and propagate the signals that we view as relevant —multiplying their impact in our memory a hundred fold. My second pillar is active engagement: a passive organism learns almost nothing, because learning requires an active generation of hypotheses, with motivation and curiosity. The third pillar, and the flip side to active engagement, is error feedback: whenever we are surprised because the world violates our expectations, error signals spread throughout our brain. They correct our mental models, eliminate inappropriate hypotheses, and stabilize the most accurate ones. Finally, the fourth pillar is consolidation: over time, our brain compiles what it has acquired and transfers it into long-term memory, thus freeing neural resources for further learning. Repetition plays an essential role in this consolidation process. Even sleep, far from being a period of inactivity, is a privileged moment during which the brain revisits its past states, at a faster pace, and recodes the knowledge acquired during the day.

These four pillars are universal: babies, children, and adults of all ages continually deploy them whenever they exercise their ability to learn. This is why we should all learn to master them—it is how we can learn to learn. In the conclusion, I will come back to the practical consequences of these scientific advances. Changing our practices

Copyrighted material

at school, at home, or at work is not necessarily as complicated as we think. Very simple ideas about play, curiosity, socialization, concentration, and sleep can augment what is already our brain's greatest talent: learning.

Copyrighted material

CHAPTER 1
Seven Definitions of Learning

WHAT DOES “LEARNING” MEAN? MY FIRST AND MOST GENERAL DEFINITION is the following: to learn is to form an internal model of the external world.

You may not be aware of it, but your brain has acquired thousands of internal models of the outside world. Metaphorically speaking, they are like miniature mock-ups more or less faithful to the reality they represent. We all have in our brains, for example, a mental map of our neighborhood and our home—all we have to do is close our eyes and envision them with our thoughts. Obviously, none of us were born with this mental map—we had to acquire it through learning.

The richness of these mental models, which are, for the most part, unconscious, exceeds our imagination. For example, you possess a vast mental model of the English language, which allows you to understand the words you are reading right now and guess that *plastovski* is not an English word, whereas *swoon* and *wistful* are, and *dragostan* could be. Your brain also includes several models of your body: it constantly uses them to map the position of your limbs and to direct them while maintaining your balance. Other mental models encode your knowledge of objects and your interactions with them: knowing how to hold a pen, write, or ride a bike. Others even represent the minds of others: you possess a vast mental catalog of people who are close to you, their appearances, their voices, their tastes, and their quirks.

These mental models can generate hyper-realistic simulations of the universe around us. Did you ever notice that your brain sometimes projects the most authentic virtual reality shows, in which you can walk, move, dance, visit new places, have brilliant conversations, or feel strong emotions? These are your dreams! It is fascinating to realize that all the thoughts that come to us in our dreams, however complex, are simply the product of our free-running internal models of the world.

But we also dream up reality when awake: our brain constantly projects hypotheses and interpretative frameworks on the outside world. This is because, unbeknownst to us, every image that appears on our retina is ambiguous—whenever we see a plate, for instance, the image is compatible with an infinite number of ellipses. If we see the plate as round, even though the raw sense data picture it as an oval, it is because our brain supplies additional data: it has learned that the round shape is the most likely interpretation. Behind the scenes, our sensory areas ceaselessly compute with probabilities, and only the most likely model makes it into our consciousness. It is the brain’s projections that ultimately give meaning to the flow of data that reaches us

Copyrighted material

from our senses. In the absence of an internal model, raw sensory inputs would remain meaningless.

Learning allows our brain to grasp a fragment of reality that it had previously missed and to use it to build a new model of the world. It can be a part of external reality, as when we learn history, botany, or the map of a city, but our brain also learns to map the reality internal to our bodies, as when we learn to coordinate our actions and concentrate our thoughts in order to play the violin. In both cases, our brain *internalizes* a new aspect of reality: it adjusts its circuits to appropriate a domain that it had not mastered before.

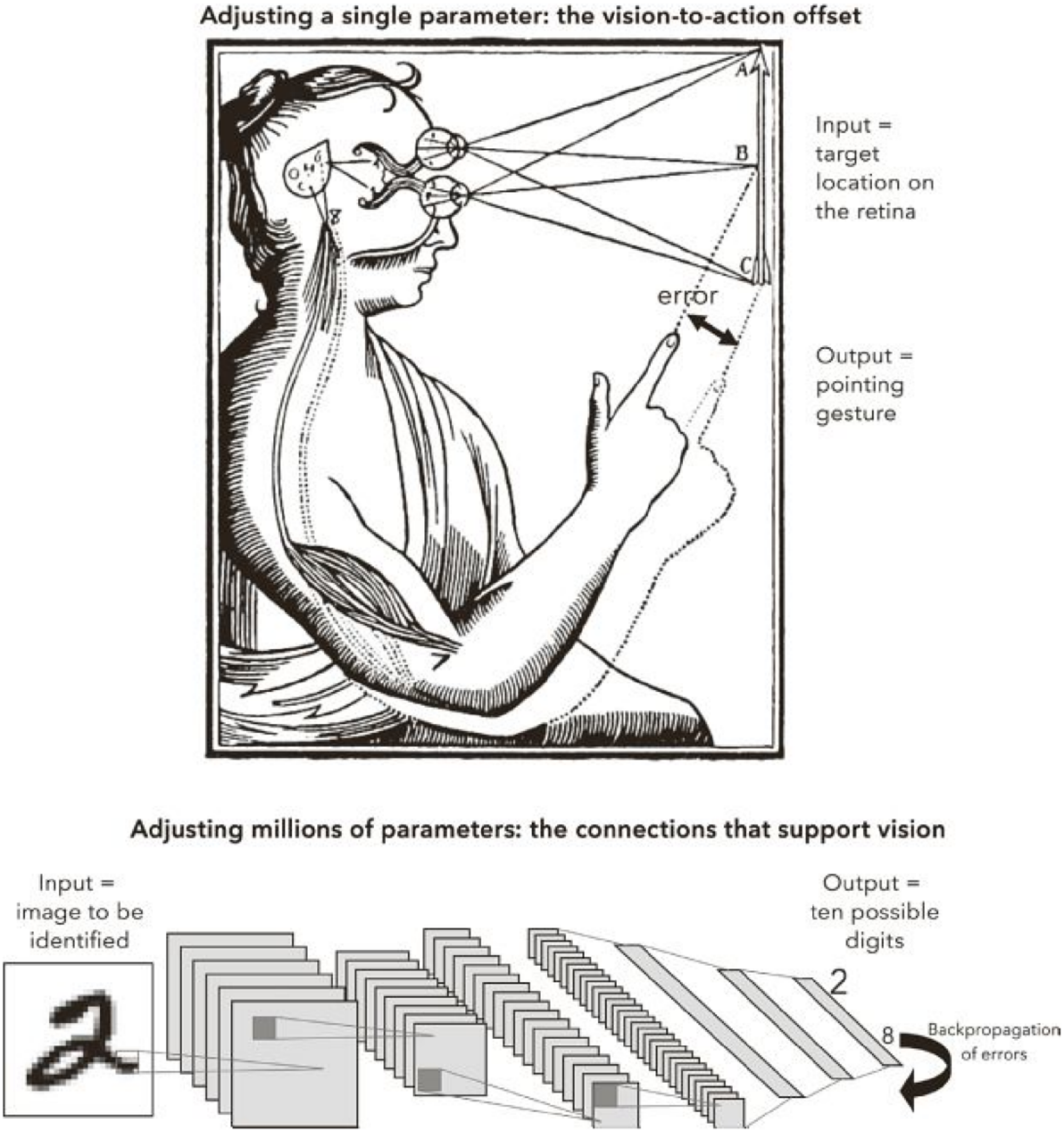
Such adjustments, of course, have to be pretty clever. The power of learning lies in its ability to adjust to the external world and to correct for errors—but how does the brain of the learner "know" how to update its internal model when, say, it gets lost in its neighborhood, falls from its bike, loses a game of chess, or misspells the word *ecstasy*? We will now review seven key ideas that lie at the heart of present-day machine-learning algorithms and that may apply equally well to our brains—seven different definitions of what "learning" means.

LEARNING IS ADJUSTING THE PARAMETERS OF A MENTAL MODEL

Adjusting a mental model is sometimes very simple. How, for example, do we reach out to an object that we see? In the seventeenth century, René Descartes (1596–1650) had already guessed that our nervous system must contain processing loops that transform visual inputs into muscular commands (see the figure on the next page). You can experience this for yourself: try grabbing an object while wearing somebody else's glasses, preferably someone who is very nearsighted. Even better, if you can, get a hold of prisms that shift your vision a dozen degrees to the left and try to catch the object.[1] You will see that your first attempt is completely off: because of the prisms, your hand reaches to the right of the object that you are aiming for. Gradually, you adjust your movements to the left. Through successive trial and error, your gestures become more and more precise, as your brain learns to correct the offset of your eyes. Now take off the glasses and grab the object: you'll be surprised to see that your hand goes to the wrong location, now way too far to the left!

So, what happened? During this brief learning period, your brain adjusted its internal model of vision. A parameter of this model, one that corresponds to the offset between the visual scene and the orientation of your body, was set to a new value. During this recalibration process, which works by trial and error, what your brain did can be likened to what a hunter does in order to adjust his rifle's viewfinder: he takes a test shot, then uses it to adjust his scope, thus progressively shooting more and more accurately. This type of learning can be very fast: a few trials are enough to correct the gap between vision and action. However, the new parameter setting is not compatible with the old one—hence the systematic error we all make when we remove the prisms and return to normal vision.

Copyrighted material



Adjusting a single parameter: the vision-to-action offset

A
B
C
Input = target location on the retina
error
Output = pointing gesture

Adjusting millions of parameters: the connections that support vision

Input = image to be identified
Output = ten possible digits
2
8
Backpropagation of errors

What is learning? To learn is to adjust the parameters of an internal model. Learning to aim with one's finger, for example, consists of setting the offset between vision and action: each aiming error provides useful information that allows one to reduce the gap. In artificial neural networks, although the number of settings is much larger, the logic is the same. Recognizing a character requires the fine-tuning of millions of connections. Again, each error—here, the incorrect activation of the output "8"—can be back-propagated and used to adjust the values of the connections, thus improving performance on the next test.

Undeniably, this type of learning is a little particular, because it requires the adjustment of only a single parameter (viewing angle). Most of our learning is much more elaborate and requires adjusting tens, hundreds, or even thousands of millions of parameters (every synapse in the relevant brain circuit). The principle, however, is always the same: it boils down to searching, among myriad possible settings of the internal model, for those that best correspond to the state of the external world.

An infant is born in Tokyo. Over the next two or three years, its internal model of language will have to adjust to the characteristics of the Japanese language. This baby's brain is like a machine with millions of settings at each level. Some of these settings, at

Copyrighted material

the auditory level, determine which inventory of consonants and vowels is used in Japanese and the rules that allow them to be combined. A baby born into a Japanese family must discover which phonemes make up Japanese words and where to place the boundaries between those sounds. One of the parameters, for example, concerns the distinction between the sounds /R/ and /L/: this is a crucial contrast in English, but not in Japanese, which makes no distinction between Bill Clinton's election and his erection .... Each baby must thus fix a set of parameters that collectively specify which categories of speech sounds are relevant for his or her native language.

A similar learning procedure is duplicated at each level, from sound patterns to vocabulary, grammar, and meaning. The brain is organized as a hierarchy of models of reality, each nested inside the next like Russian dolls—and learning means using the incoming data to set the parameters at every level of this hierarchy. Let's consider a high-level example: the acquisition of grammatical rules. Another key difference which the baby must learn, between Japanese and English, concerns the order of words. In a canonical sentence with a subject, a verb, and a direct object, the English language first states the subject, then the verb, and finally its object: "John + eats + an apple." In Japanese, on the other hand, the most common order is subject, then object, then verb: "John + an apple + eats." What is remarkable is that the order is also reversed for prepositions (which logically become post-positions), possessives, and many other parts of speech. The sentence "My uncle wants to work in Boston," thus becomes mumbo jumbo worthy of Yoda from Star Wars: "Uncle my, Boston in, work wants"—which makes perfect sense to a Japanese speaker.

Fascinatingly, these reversals are not independent of one another. Linguists think that they arise from the setting of a single parameter called the "head position": the defining word of a phrase, its head, is always placed first in English (in Paris, my uncle, wants to live), but last in Japanese (Paris in, uncle my, live wants). This binary parameter distinguishes many languages, even some that are not historically linked (the Navajo language, for example, follows the same rules as Japanese). In order to learn English or Japanese, one of the things that a child must figure out is how to set the head position parameter in his internal language model.

LEARNING IS EXPLOITING A COMBINATORIAL EXPLOSION

Can language learning really be reduced to the setting of some parameters? If this seems hard to believe, it is because we are unable to fathom the extraordinary number of possibilities that open up as soon as we increase the number of adjustable parameters. This is called the "combinatorial explosion"—the exponential increase that occurs when you combine even a small number of possibilities. Suppose that the grammar of the world's languages can be described by about fifty binary parameters, as some linguists postulate. This yields 2^{50} combinations, which are over one million billion possible languages, or 1 followed by fifteen zeros! The syntactic rules of the world's three thousand languages easily fit into this gigantic space. However, in our brain, there aren't just fifty adjustable parameters, but an astoundingly larger number: eighty-six billion neurons, each with about ten thousand synaptic contacts whose strength can vary. The space of mental representations that opens up is practically infinite.

Human languages heavily exploit these combinations at all levels. Consider, for instance, the mental lexicon: the set of words that we know and whose model we carry around with us. Each of us has learned about fifty thousand words with the most diverse meanings. This seems like a huge lexicon, but we manage to acquire it in about

Copyrighted material

a decade because we can decompose the learning problem. Indeed, considering that these fifty thousand words are on average two syllables, each consisting of about three phonemes, taken from the forty-four phonemes in English, the binary coding of all these words requires less than two million elementary binary choices (“bits,” whose value is 0 or 1). In other words, all our knowledge of the dictionary would fit in a small 250-kilobyte computer file (each byte comprising eight bits).

This mental lexicon could be compressed to an even smaller size if we took into account the many redundancies that govern words. Drawing six letters at random, like “xfdrga,” does not generate an English word. Real words are composed of a pyramid of syllables that are assembled according to strict rules. And this is true at all levels: sentences are regular collections of words, which are regular collections of syllables, which are regular collections of phonemes. The combinations are both vast (because one chooses among several tens or hundreds of elements) and bounded (because only certain combinations are allowed). To learn a language is to discover the parameters that govern these combinations at all levels.

In summary, the human brain breaks down the problem of learning by creating a hierarchical, multilevel model. This is particularly obvious in the case of language, from elementary sounds to the whole sentence or even discourse—but the same principle of hierarchical decomposition is reproduced in all sensory systems. Some brain areas capture low-level patterns: they see the world through a very small temporal and spatial window, thus analyzing the smallest patterns. For example, in the primary visual area, the first region of the cortex to receive visual inputs, each neuron analyzes only a very small portion of the retina. It sees the world through a pinhole and, as a result, discovers very low-level regularities, such as the presence of a moving oblique line. Millions of neurons do the same work at different points in the retina, and their outputs become the inputs of the next level, which thus detects “regularities of regularities,” and so on and so forth. At each level, the scale broadens: the brain seeks regularities on increasingly vast scales, in both time and space. From this hierarchy emerges the ability to detect increasingly complex objects or concepts: a line, a finger, a hand, an arm, a human body ... no, wait, two, there are two people facing each other, a handshake .... It is the first Trump-Macron encounter!

LEARNING IS MINIMIZING ERRORS

The computer algorithms that we call “artificial neural networks” are directly inspired by the hierarchical organization of the cortex. Like the cortex, they contain a pyramid of successive layers, each of which attempts to discover deeper regularities than the previous one. Because these consecutive layers organize the incoming data in deeper and deeper ways, they are also called “deep networks.” Each layer, by itself, is capable of discovering only an extremely simple part of the external reality (mathematicians speak of a linearly separable problem, i.e., each neuron can separate that data into only two categories, A and B, by drawing a straight line through them). Assemble many of these layers, however, and you get an extremely powerful learning device, capable of discovering complex structures and adjusting to very diverse problems. Today’s artificial neural networks, which take advantage of the advances in computer chips, are also deep, in the sense that they contain dozens of successive layers. These layers become increasingly insightful and capable of identifying abstract properties the further away they are from the sensory input.

Let’s take the example of the LeNet algorithm, created by the French pioneer of neural networks, Yann LeCun (see figure 2 in the color insert).[2] As early as the 1990s,

Copyrighted material

One of the problems with the error correction procedure I just described is that it can get stuck on a set of parameters that is not the best. Imagine a golf ball rolling on the green, always along the line of the steepest slope: it may get stuck in a small depression in the ground, preventing it from reaching the lowest point of the whole landscape, the absolute optimum. Similarly, the gradient descent algorithm sometimes gets stuck at a point that it cannot exit. This is called a "local minimum": a well in parameter space, a trap from which the learning algorithm cannot escape because it seems impossible to do better. At this moment, learning gets stuck, because all changes seem counterproductive: each of them increases the error rate. The system feels that it has learned all it can. It remains blind to the presence of much better settings, perhaps only a few steps away in parameter space. The gradient descent algorithm does not "see" them because it refuses to go up the hump in order to go back down the other side of the dip. Shortsighted, it ventures only a small distance from its starting point and may therefore miss out on better but distant configurations.

Does the problem seem too abstract to you? Think about a concrete situation: You go shopping at a food market, where you spend some time looking for the cheapest products. You walk down an aisle, pass the first seller (who seems overpriced), avoid the second (who is always very expensive), and finally stop at the third stand, which seems much cheaper than the previous ones. But who's to say that one aisle over, or perhaps even in the next town, the prices would not be even more enticing? Focusing on the best *local* price does not guarantee finding the *global* minimum.

Frequently confronted with this difficulty, computer scientists employ a panoply of tricks. Most of them consist of introducing a bit of randomness in the search for the best parameters. The idea is simple: instead of looking in only one aisle of the market, take a step at random; and instead of letting the golf ball roll gently down the slope, give it a shake, thus reducing its chance of getting stuck in a trough. On occasion, stochastic search algorithms try a distant and partially random setting, so that if a better solution is within reach, they have a chance of finding it. In practice, one can introduce some degree of randomness in various ways: setting or updating the parameters at random, diversifying the order of the examples, adding some noise to the data, or using only a random fraction of the connections—all these ideas improve the robustness of learning.

Some machine learning algorithms also get their inspiration from the Darwinian algorithm that governs the evolution of species: during parameter optimization, they introduce mutations and random crossings of previously discovered solutions. As in biology, the rate of these mutations must be carefully controlled in order to explore new solutions without wasting too much time in hazardous attempts.

Another algorithm is inspired by blacksmith forges, where craftspeople have learned to optimize the properties of metal by "annealing" it. Applied when one wants to forge an exceptionally strong sword, the method of annealing consists of heating the metal several times, at lower and lower temperatures, to increase the chance that the atoms arrange themselves in a regular configuration. The process has now been transposed to computer science: the simulated annealing algorithm introduces random changes in the parameters, but with a virtual "temperature" that gradually decreases. The probability of a chance event is high at the beginning but steadily declines until the system is frozen in an optimal setting.

Computer scientists have found all these tricks to be remarkably effective—so perhaps it should be no surprise that, in the course of evolution, some of them were internalized in our brains. Random exploration, stochastic curiosity, and noisy

Copyrighted material

neuronal firing all play an essential role in learning for *Homo sapiens.* Whether we are playing rock, paper, scissors; improvising on a jazz theme; or exploring the possible solutions to a math problem, randomness is an essential ingredient of a solution. As we shall see, whenever children go into learning mode—that is, when they play—they explore dozens of possibilities with a good dose of randomness. And during the night, their brains continue juggling ideas until they hit upon one that best explains what they experienced during the day. In the third section of this book, I will come back to what we know about the semi-random algorithm that governs the extraordinary curiosity of children—and the rare adults who have managed to keep a child's mind.

LEARNING IS OPTIMIZING A REWARD FUNCTION

Remember LeCun's LeNet system, which recognizes the shapes of numbers? In order to learn, this type of artificial neural network needs to be provided with the correct answers. For each input image, it needs to know which of the ten possible numbers it corresponds to. The network can correct itself only by calculating the difference between its response and the correct answer. This procedure is known as "supervised learning": a supervisor, outside the system, knows the solution and tries to teach it to the machine. This is effective, but it should be noted that this situation, where the right answer is known in advance, is rather rare. When children learn to walk, no one tells them exactly which muscles to contract—they are simply encouraged again and again until they no longer fall. Babies learn solely on the basis of an evaluation of the result: I fell, or, on the contrary, I finally managed to walk across the room.

Artificial intelligence faces the same "unsupervised learning" problem. When a machine learns to play a video game, for example, the only thing it is told is that it must try to attain the highest score. No one tells it in advance what specific actions need to be taken to achieve this. How can it quickly find out for itself the right way of going about it?

Scientists have responded to this challenge by inventing "reinforcement learning," whereby we do not provide the system with any detail about what it must do (nobody knows!), but only with a "reward," an evaluation in the form of a quantitative score.[5] Even worse, the machine may receive its score after a delay, long after the decisive actions that led to it. Such delayed reinforcement learning is the principle by which the company DeepMind, a Google subsidiary, created a machine capable of playing chess, checkers, and Go. The problem is colossal for a simple reason: it is only at the very end that the system receives a single reward signal, indicating whether the game was won or lost. During the game itself, the system receives no feedback whatsoever—only the final checkmate counts. How, then, can the system figure out what to do at any given time? And, once the final score is known, how can the machine retrospectively evaluate its decisions?

The trick that computer scientists have found is to program the machine to do two things at the same time: to act and to self-evaluate. One half of the system, called the "critic," learns to predict the final score. The goal of this network of artificial neurons is to evaluate, as accurately as possible, the state of the game, in order to predict the final reward: Am I winning or losing? Is my balance stable, or am I about to fall? Thanks to this critic that emerges in this half of the machine, the system can evaluate its actions at every moment and not just at the end. The other half of the machine, the actor, can then use this evaluation to correct itself: Wait! I'd better avoid this or that action, because the critic thinks it will increase my chances of losing.

Copyrighted material

Trial after trial, the actor and the critic progress together: one learns to act wisely, focusing on the most effective actions, while the other learns to evaluate, ever more sharply, the consequences of these acts. In the end, unlike the famed guy who is falling from a skyscraper and exclaims, “So far, so good,” the actor-critic network becomes endowed with a remarkable prescience: the ability to predict, within the vast seas of not-yet-lost games, those that are likely to be won and those that will lead only to disaster.

The actor-critic combination is one of the most effective strategies of contemporary artificial intelligence. When backed by a hierarchical neural network, it works wonders. As early as the 1980s, it enabled a neural network to win the backgammon world cup. More recently, it enabled DeepMind to create a multifunctional neural network capable of learning to play all kinds of video games such as *Super Mario* and *Tetris*.[6] One simply gives this system the pixels of the image as an input, the possible actions as an output, and the score of the game as a reward function. The machine learns everything else. When it plays *Tetris*, it discovers that the screen is made up of shapes, that the falling one is more important than the others, that various actions can change its orientation and its position, and so on and so forth—until the machine turns into an artificial player of formidable effectiveness. And when it plays *Super Mario*, the change in inputs and rewards teaches it to attend to completely different settings: what pixels form Mario’s body, how he moves, where the enemies are, the shapes of walls, doors, traps, bonuses ... and how to act in front of each of them. By adjusting its parameters, i.e., the millions of connections that link the layers together, a single network can adapt to all kinds of games and learn to recognize the shapes of *Tetris*, *Pac-Man*, or *Sonic the Hedgehog*.

What is the point of teaching a machine to play video games? Two years later, DeepMind engineers used what they had learned from game playing to solve an economic problem of vital interest: How should Google optimize the management of its computer servers? The artificial neural network remained similar; the only things that changed were the inputs (date, time, weather, international events, search requests, number of people connected to each server, etc.), the outputs (turn on or off this or that server on various continents), and the reward function (consume less energy). The result was an instant drop in power consumption. Google reduced its energy bill by up to 40 percent and saved tens of millions of dollars—even after myriad specialized engineers had already tried to optimize those very servers. Artificial intelligence has truly reached levels of success that can turn whole industries upside down.

DeepMind has achieved even more amazing feats. As everyone probably knows, its AlphaGo program managed to beat eighteen-time world champion Lee Sedol in the game of Go, considered until very recently the Everest of artificial intelligence.[7] This game is played on a vast square checkerboard (a *goban*) with nineteen positions on each side, for a total of 361 places where black and white pieces can be played. The number of combinations is so vast that it is strictly impossible to systematically explore all the future moves available to each player. And yet reinforcement learning allowed the AlphaGo software to recognize favorable and unfavorable combinations better than any human player. One of the many tricks was to make the system play against itself, just as a chess player trains by playing both white and black. The idea is simple: at the end of each game, the winning software strengthens its actions, while the loser weakens them—but both have also learned to evaluate their moves more efficiently.

Copyrighted material

We happily mock Baron Munchausen, who, in his fabled *Adventures*, foolishly attempts to fly away by pulling on his bootstraps. In artificial intelligence, however, Munchausen's mad method gave birth to a rather sophisticated strategy, aptly called "bootstrapping"—little by little, starting from a meaningless architecture devoid of knowledge, a neural network can become a world champion, simply by playing against itself.

This idea of increasing the speed of learning by letting two networks collaborate—or, on the contrary, compete—continues to lead to major advances in artificial intelligence. One of the most recent ideas, called "adversarial learning,"[8] consists of training two opponent systems: one that learns to become an expert (say, in Van Gogh's paintings) and another whose sole goal is to make the first one fail (by learning to become a brilliant forger of false Van Goghs). The first system gets a bonus whenever it successfully identifies a genuine Van Gogh painting, while the second is rewarded whenever it manages to fool the other's expert eye. This adversarial learning algorithm yields not just one but two artificial intelligences: a world authority in Van Gogh, fond of the smallest details that can authenticate a true painting by the master, and a genius forger, capable of producing paintings that can fool the best of experts. This sort of training can be likened to the preparation for a presidential debate: a candidate can sharpen her training by hiring someone to imitate her opponent's best lines.

Could this approach apply to a single human brain? Our two hemispheres and numerous subcortical nuclei also host a whole collection of experts who fight, coordinate, and evaluate one another. Some of the areas in our brain learn to simulate what others are doing; they allow us to foresee and imagine the results of our actions, sometimes with a realism worthy of the best counterfeiters: our memory and imagination can make us see the seaside bay where we swam last summer, or the door handle that we grab in the dark. Some areas learn to criticize others: they constantly assess our abilities and predict the rewards or punishments we might get. These are the areas that push us to act or to remain silent. We will also see that metacognition—the ability to know oneself, to self-evaluate, to mentally simulate what would happen if we acted this way or that way—plays a fundamental role in human learning. The opinions we form of ourselves help us progress or, in some cases, lock us into a vicious circle of failure. Thus, it is not inappropriate to think of the brain as a collection of experts that collaborate and compete.

LEARNING IS RESTRICTING SEARCH SPACE

Contemporary artificial intelligence still faces a major problem. The more parameters the internal model has, the more difficult it is to find the best way to adjust it. And in current neural networks, the search space is immense. Computer scientists therefore have to deal with a massive combinatorial explosion: at each stage, millions of choices are available, and their combinations are so vast that it is impossible to explore them all. As a result, learning is sometimes exceedingly slow: it takes billions of attempts to move the system in the right direction within this immense landscape of possibilities. And the data, however large, become scarce relative to the gigantic size of that space. This issue is called the "curse of dimensionality"—learning can become very hard when you have millions of potential levers to pull.

The immense number of parameters that neural networks possess often leads to a second obstacle, which is called "overfitting" or "overlearning": the system has so

Copyrighted material

many degrees of freedom that it finds it easier to memorize all the details of each example than it is to identify a more general rule that can explain them.

As John von Neumann (1903–57), the father of computer science, famously said, "With four parameters I can fit an elephant, and with five I can make him wiggle his trunk." What he meant is that having too many free parameters can be a curse: it's all too easy to "overfit" any data simply by memorizing every detail, but that does not mean that the resulting system captures anything significant. You can fit the pachyderm's profile without understanding anything deep about elephants as a species. Having too many free parameters can be detrimental to abstraction. While the system easily learns, it is unable to generalize to new situations. Yet this ability to generalize is the key to learning. What would be the point of a machine that could recognize a picture that it has already seen, or win a game of Go that it has already played? Obviously, the real aim is to recognize any picture, or to win against any player, whether the circumstances are familiar or new.

Again, computer scientists are investigating various solutions to these problems. One of the most effective interventions, which can both accelerate learning and improve generalization, is to simplify the model. When the number of parameters to be adjusted is minimized, the system can be forced to find a more general solution. This is the key insight that led LeCun to invent *convolutional neural networks*, an artificial learning device which has become ubiquitous in the field of image recognition.[9] The idea is simple: in order to recognize the items in a picture, you pretty much have to do the same job everywhere. In a photo, for example, faces may appear anywhere. To recognize them, one should apply the same algorithm to every part of the picture (e.g., to look for an oval, a pair of eyes, etc.). There is no need to learn a different model at each point of the retina: what is learned in one place can be reused everywhere else.

Over the course of learning, LeCun's convolutional neural networks apply whatever they learn from a given region to the entire network, at all levels and on ever wider scales. They therefore have a much smaller number of parameters to learn: by and large, the system has to tune only a single filter that it applies everywhere, rather than a plethora of different connections for each location in the image. This simple trick massively improves performance, especially generalization to new images. The reason is simple: the algorithm that runs on a new image benefits from the immense experience it gained from every point of every photo that it has ever seen. It also speeds up learning, since the machine explores only a subset of vision models. Prior to learning, it already knows something important about the world: that the same object can appear anywhere in the image.

This trick generalizes to many other domains. To recognize speech, for example, one must abstract away from the specifics of the speaker's voice. This is achieved by forcing a neural network to use the same connections in different frequency bands, whether the voice is low or high. Reducing the number of parameters that must be adjusted leads to greater speeds and better generalization to new voices: the advantage is twofold, and this is how your smartphone is able to respond to your voice.

LEARNING IS PROJECTING A PRIORI HYPOTHESES

Yann LeCun's strategy provides a good example of a much more general notion: the exploitation of innate knowledge. Convolutional neural networks learn better and faster than other types of neural networks because they do not learn everything. They

Copyrighted material

CHAPTER 2

Why Our Brain Learns Better Than Current Machines

THE RECENT SURGE OF PROGRESS IN ARTIFICIAL INTELLIGENCE MAY SUGGEST that we have finally discovered how to copy and even surpass human learning and intelligence. According to some self-proclaimed prophets, machines are about to overtake us. Nothing could be further from the truth. In fact, most cognitive scientists, while admiring recent advances in artificial neural networks, are well aware of the fact that these machines remain highly limited. In truth, most artificial neural networks implement only the operations that our brain performs unconsciously, in a few tenths of a second, when it perceives an image, recognizes it, categorizes it, and accesses its meaning.[1] However, our brain goes much further: it is able to explore the image consciously, carefully, step by step, for several seconds. It formulates symbolic representations and explicit theories of the world that we can share with others through language.

Operations of this nature—slow, reasoned, symbolic—remain (for now) the exclusive privilege of our species. Current machine learning algorithms capture them poorly. Although there is constant progress in the fields of machine translation and logical reasoning, a common criticism of artificial neural networks is that they attempt to learn everything at the same level, as if every problem were a matter of automatic classification. To a man with a hammer, everything looks like a nail! But our brain is much more flexible. It quickly manages to prioritize information and, whenever possible, extract general, logical, and explicit principles.

WHAT IS ARTIFICIAL INTELLIGENCE MISSING?

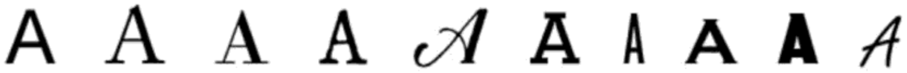
It is interesting to try to clarify what artificial intelligence is still missing, because this is also a way to identify what is unique about our species' learning abilities. Here is a short and probably still partial list of functions that even a baby possesses and that most current artificial systems are missing:

> **Learning abstract concepts.** Most artificial neural networks capture only the very first stages of information processing—those that, in less than a fifth of a second, parse an image in the visual areas of our brain. Deep learning algorithms are far from being as deep as some people claim. According to Yoshua Bengio, one of the inventors of deep learning algorithms, they actually tend to learn superficial statistical regularities in data, rather than high-level

Copyrighted material

abstract concepts.[2] To recognize an object, for instance, they often rely on the presence of a few shallow features in the image, such as a specific color or shape. Change these details and their performance collapses: contemporary convolutional neural networks are unable to recognize what constitutes the essence of an object; they have difficulty understanding that a chair remains a chair whether it has four legs or just one, and whether it is made of glass, metal, or inflatable plastic. This inclination to attend to superficial features makes these networks susceptible to massive errors. There is a whole literature on how to fool a neural network: take a banana and modify a few pixels or put a particular sticker on it, and the neural network will think it's a toaster!

True enough, when you flash an image to a person for a split second, they will sometimes make the same kinds of errors as a machine and may mistake a dog for a cat.[3] However, as soon as humans are given a little more time, they correct their errors. Unlike a computer, we possess the ability to question our beliefs and refocus our attention on those aspects of an image that do not fit with our first impression. This second analysis, conscious and intelligent, calls upon our general powers of reasoning and abstraction. Artificial neural networks neglect an essential point: human learning is not just the setting of a pattern-recognition filter, but the forming of an abstract model of the world. By learning to read, for example, we have acquired an abstract concept of each letter of the alphabet, which allows us to recognize it in all its disguises, as well as generate new versions:



The cognitive scientist Douglas Hofstadter once said that the real challenge for artificial intelligence was to recognize the letter A! This quip was undoubtedly an exaggeration, but a profound one nevertheless: even in this most trivial context, humans deploy an unmatched knack for abstraction. This feat is at the origin of an amusing occurrence of daily life: the CAPTCHA, the little chain of letters that some websites ask you to recognize in order to prove you are a human being, not a machine. For years, CAPTCHAs have withstood machines. But computer science is evolving fast: in 2017, an artificial system managed to recognize CAPTCHAs at an almost humanlike level.[4] Unsurprisingly, this algorithm mimics the human brain in several respects. A genuine tour de force, it manages to extract the skeleton of each letter, the inner essence of the letter A, and uses all the resources of statistical reasoning to verify whether this abstract idea applies to the current image. Yet this computer algorithm, however sophisticated, applies only to CAPTCHAs. Our brains apply this ability for abstraction to all aspects of our daily lives.

Data-efficient learning. Everyone agrees that today's neural networks learn far too slowly: they need thousands, millions, even billions of data points to develop an intuition of a domain. We even have experimental evidence of this sluggishness. For instance, it takes no less than nine hundred hours of play for the neural network designed by DeepMind to reach a reasonable level on an Atari console—while a human being reaches the same level in two hours![5] Another example is language learning. Psycholinguist Emmanuel Dupoux estimates that in most French families, children hear about five hundred to one thousand hours of speech per year, which is more than enough for them to

Copyrighted material

PENGUIN BOOKS

UK | USA | Canada | Ireland | Australia
India | New Zealand | South Africa

Penguin Books is part of the Penguin Random House group of companies whose addresses can be found at global.penguinrandomhouse.com.

Penguin
Random House
UK

First published in the United States of America by Viking, an imprint of Penguin Random House LLC 2020
First published in Great Britain by Allen Lane 2020
Published in Penguin Books 2021

Copyright © Stanislas Dehaene, 2020

Based on, in part, *Apprendre!*, published in France by Odile Jacob, Paris, in 2018

Pages 321–324 constitute an extension to the copyright page

The moral right of the author has been asserted

Cover illustration by Théo

ISBN: 978-0-241-36647-9

This ebook is copyright material and must not be copied, reproduced, transferred, distributed, leased, licensed or publicly performed or used in any way except as specifically permitted in writing by the publishers, as allowed under the terms and conditions under which it was purchased or as strictly permitted by applicable copyright law. Any unauthorized distribution or use of this text may be a direct infringement of the author's and publisher's rights and those responsible may be liable in law accordingly.

Copyrighted material