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*.

STANISLAS DEHAENE

How We Learn

The New Science of Education and the Brain

PENGUIN BOOKS

CONTENTS

INTRODUCTION

IN SEPTEMBER 2009, AN EXTRAORDINARY CHILD FORCED ME TO DRASTI-cally 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

> At its core, intelligence can be viewed as a process that converts unstructured information into useful and actionable knowledge.
>
> Demis Hassabis,
> founder of the AI company DeepMind (2017)

What is learning? In many Latin languages, *learning* has the same root as *apprehending*: *apprendre* in French, *aprender* in Spanish and Portuguese. . . . Indeed, learning is grasping a fragment of reality, catching it, and bringing it inside our brains. In cognitive science, we say that learning consists of forming an internal model of the world. Through learning, the raw data that strikes our senses turns into refined ideas, abstract enough to be reused in a new context—smaller-scale models of reality.

In the following pages, we will review what artificial intelligence and cognitive science have taught us about how such internal models emerge, in both brains and machines. How does the representation of information change when we learn? How can we understand it at a level that is common to any organism, human, animal, or machine? By reviewing the various tricks that engineers have designed to allow machines to learn, we will progressively conjure up a sharper picture of the amazing computations that infants must

perform as they learn to see, speak, and write. In fact, as we shall see, the infant brain keeps the upper hand: despite their successes, current learning algorithms capture only a fraction of the abilities of the human brain. Understanding exactly where the machine learning metaphor breaks down, and where even an infant's brain still surpasses the most powerful computer, we will delineate exactly what "learning" means.

CHAPTER 1

Seven Definitions of Learning

WHAT DOES "LEARNING" MEAN? MY FIRST AND MOST GENERAL DEFINI- tion 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 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

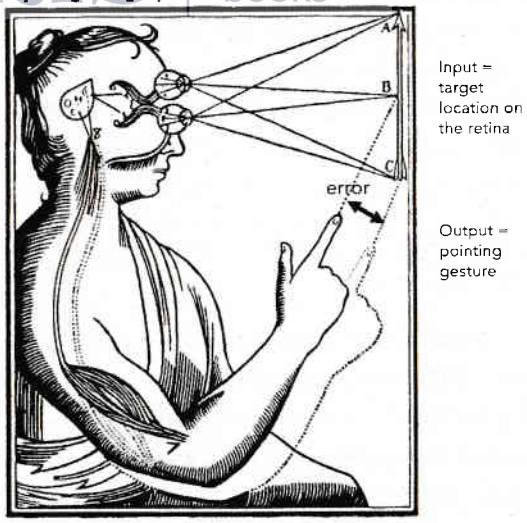
when, say, it gets lost in its neighborhood, falls from its bike, loses a game of chess, or misspells the word *ecstasy*? We will now review seven key ideas that lie at the heart of present-day machine-learning algorithms and that may apply equally well to our brains—seven different definitions of what "learning" means.

LEARNING IS ADJUSTING THE PARAMETERS OF A MENTAL MODEL

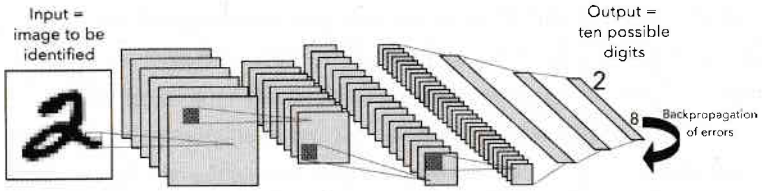
Adjusting a mental model is sometimes very simple. How, for example, do we reach out to an object that we see? In the seventeenth century, René Descartes (1596–1650) had already guessed that our nervous system must contain processing loops that transform visual inputs into muscular commands (see the 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

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



Adjusting millions of parameters: the connections that support vision



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.

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.

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 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 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, this neural network achieved remarkable performance in the recognition of handwritten characters. For years, Canada Post used it to automatically process handwritten postal codes. How does it work? The algorithm receives the image of a written character as an input, in the form of pixels, and it proposes, as an output, a tentative interpretation: one out of the ten possible digits or twenty-six letters. The artificial network contains a hierarchy of processing units that look a bit like neurons and form successive layers. The first layers are connected directly with the image: they apply simple filters that recognize lines and curve fragments. The layers higher up in the hierarchy, however, contain wider and more complex filters. Higher-level units can therefore learn to recognize larger and larger portions of the image: the curve of a 2, the loop of an O, or the parallel lines of a Z . . . until we reach, at the output level, artificial neurons that respond to a character regardless of its position, font, or

case. All these properties are not imposed by a programmer: they result entirely from the millions of connections that link the units. These connections, once adjusted by an automated algorithm, define the filter that each neuron applies to its inputs: their settings explain why one neuron responds to the number 2 and another to the number 3.

How are these millions of connections adjusted? Just as in the case of prism glasses! On each trial, the network gives a tentative answer, is told whether it made an error, and adjusts its parameters to try to reduce this error on the next trial. Every wrong answer provides valuable information. With its sign (like a gesture too far to the right or too far to the left), the error tells the system what it should have done in order to succeed. By going back to the source of the error, the machine discovers how the parameters should have been set to avoid the mistake.

Let's revisit the example of the hunter adjusting his rifle's scope. The learning procedure is elementary. The hunter shoots and finds he's aimed five centimeters too far to the right. He now has essential information, both on the amplitude (five centimeters) and on the sign of the error (too far to the right). This information allows him to correct his shot. If he is a bit clever, he can infer in which direction to make the correction: if the bullet has deflected to the right, he should shift the scope one hair to the left. Even if he's not that astute, he can casually try a different aim and test whether, if he turns the scope to the right, the offset increases or decreases. In this manner, through trial and error, the hunter can progressively discover which adjustment reduces the size of the gap between his intended target and his actual shot.

In modifying his sight to maximize his accuracy, our brave hunter is applying a learning algorithm without even knowing it. He is implicitly calculating what mathematicians call the "derivative," or gradient, of the system, and is using the "gradient descent algorithm": he learns to move his rifle's viewfinder in the most efficient direction, the one that reduces the probability of making a mistake.

Most artificial neural networks used in present-day artificial intelligence, despite their millions of inputs, outputs, and adjustable parameters, operate