A THOUSAND BRAINS

A NEW THEORY OF INTELLIGENCE

JEFF HAWKINS

With a foreword by Richard Dawkins

BASIC BOOKS
New York

Contents

BASIC
BOOKS

Foreword by Richard Dawkins

Don't read this book at bedtime. Not that it's frightening. It won't give you nightmares. But it is so exhilarating, so stimulating, it'll turn your mind into a whirling maelstrom of excitingly provocative ideas—you'll want to rush out and tell someone rather than go to sleep. It is a victim of this maelstrom who writes the foreword, and I expect it'll show.

Charles Darwin was unusual among scientists in having the means to work outside universities and without government research grants. Jeff Hawkins might not relish being called the Silicon Valley equivalent of a gentleman scientist but—well, you get the parallel. Darwin's powerful idea was too revolutionary to catch on when expressed as a brief article, and the Darwin-Wallace joint papers of 1858 were all but ignored. As Darwin himself said, the idea needed to be expressed at book length. Sure enough, it was his great book that shook Victorian foundations, a year later. Book-length treatment, too, is needed for Jeff Hawkins's Thousand Brains Theory. And for his notion of reference frames—"The very act of thinking is a form of movement"—bull's-eye! These two ideas are each profound enough to fill a book. But that's not all.

T. H. Huxley famously said, on closing *On the Origin of Species*, "How extremely stupid of me not to have thought of that." I'm not suggesting that brain scientists will necessarily say the same when they close this book. It is a book of many exciting ideas, rather than one huge idea like Darwin's.

I suspect that not just T. H. Huxley but his three brilliant grandsons would have loved it: Andrew because he discovered how the nerve impulse works (Hodgkin and Huxley are the Watson and Crick of the nervous system); Aldous because of his visionary and poetic voyages to the mind's furthest reaches; and Julian because he wrote this poem, extolling the brain's capacity to construct a model of reality, a microcosm of the universe:

The world of things entered your infant mind
To populate that crystal cabinet.
Within its walls the strangest partners met,
And things turned thoughts did propagate their kind.

For, once within, corporeal fact could find
A spirit. Fact and you in mutual debt
Built there your little microcosm—which yet
Had hugest tasks to its small self assigned.

Dead men can live there, and converse with stars:
Equator speaks with pole, and night with day;
Spirit dissolves the world's material bars—
A million isolations burn away.
The Universe can live and work and plan,
At last made God within the mind of man.

The brain sits in darkness, apprehending the outside world only through a hailstorm of Andrew Huxley's nerve impulses. A nerve impulse from the eye is no different from one from the ear or the big toe. It's where they end up in the brain that sorts them out. Jeff Hawkins is not the first scientist or philosopher to suggest that the reality we perceive is a constructed reality, a model, updated and informed by bulletins streaming in from the senses. But Hawkins is, I think, the first to give eloquent space to the idea that there is not one such model but thousands, one in each of the many neatly stacked columns that constitute the brain's cortex. There are about 150,000 of these columns and they are the stars of the first section of the book, along with what he calls "frames of reference." Hawkins's thesis about both of these is provocative, and it'll be interesting to see how it is received by other brain scientists: well, I suspect. Not the least fascinating of his ideas here is that the cortical columns, in their world-modeling activities, work semi-autonomously. What "we" perceive is a kind of democratic consensus from among them.

Democracy in the brain? Consensus, and even dispute? What an amazing idea. It is a major theme of the book. We human mammals are the victims of a recurrent dispute: a tussle between

the old reptilian brain, which unconsciously runs the survival machine, and the mammalian neocortex sitting in a kind of driver's seat atop it. This new mammalian brain—the cerebral cortex—thinks. It is the seat of consciousness. It is aware of past, present, and future, and it sends instructions to the old brain, which executes them.

The old brain, schooled by natural selection over millions of years when sugar was scarce and valuable for survival, says, "Cake. Want cake. Mmmm cake. Gimme." The new brain, schooled by books and doctors over mere tens of years when sugar was overplentiful, says, "No, no. Not cake. Mustn't. Please don't eat that cake." Old brain says, "Pain, pain, horrible pain, stop the pain *immediately*." New brain says, "No, no, bear the torture, don't betray your country by surrendering to it. Loyalty to country and comrades comes before even your own life."

The conflict between the old reptilian and the new mammalian brain furnishes the answer to such riddles as "Why does pain have to be so damn painful?" What, after all, is pain for? Pain is a proxy for death. It is a warning to the brain, "Don't do that again: don't tease a snake, pick up a hot ember, jump from a great height. This time it only hurt; next time it might kill you." But now a designing engineer might say what we need here is the equivalent of a painless flag in the brain. When the flag shoots up, don't repeat whatever you just did. But instead of the engineer's easy and painless flag, what we actually get is pain—often excruciating, unbearable pain. Why? What's wrong with the sensible flag?

The answer probably lies in the disputatious nature of the brain's decision-making processes: the tussle between old brain and new brain. It being too easy for the new brain to overrule the vote of the old brain, the painless flag system wouldn't work. Neither would torture.

The new brain would feel free to ignore my hypothetical flag and endure any number of bee stings or sprained ankles or torturers' thumbscrews if, for some reason, it "wanted to." The old brain, which really "cares" about surviving to pass on the genes, might "protest" in vain. Maybe natural selection, in the interests of survival, has ensured "victory" for the old brain by making pain so damn painful that the new brain cannot overrule it. As another example, if the old brain were "aware" of the betrayal of sex's

Darwinian purpose, the act of donning a condom would be unbearably painful.

Hawkins is on the side of the majority of informed scientists and philosophers who will have no truck with dualism: there is no ghost in the machine, no spooky soul so detached from hardware that it survives the hardware's death, no Cartesian theatre (Dan Dennett's term) where a colour screen displays a movie of the world to a watching self. Instead, Hawkins proposes multiple models of the world, constructed microcosms, informed and adjusted by the rain of nerve impulses pouring in from the senses. By the way, Hawkins doesn't totally rule out the long-term future possibility of escaping death by uploading your brain to a computer, but he doesn't think it would be much fun.

Among the more important of the brain's models are models of the body itself, coping, as they must, with how the body's own movement changes our perspective on the world outside the prison wall of the skull. And this is relevant to the major preoccupation of the middle section of the book, the intelligence of machines. Jeff Hawkins has great respect, as do I, for those smart people, friends of his and mine, who fear the approach of superintelligent machines to supersede us, subjugate us, or even dispose of us altogether. But Hawkins doesn't fear them, partly because the faculties that make for mastery of chess or Go are not those that can cope with the complexity of the real world. Children who can't play chess "know how liquids spill, balls roll, and dogs bark. They know how to use pencils, markers, paper, and glue. They know how to open books and that paper can rip." And they have a self-image, a body image that emplaces them in the world of physical reality and allows them to navigate effortlessly through it.

It is not that Hawkins underestimates the power of artificial intelligence and the robots of the future. On the contrary. But he thinks most present-day research is going about it the wrong way. The right way, in his view, is to understand how the brain works and to borrow its ways but hugely speed them up.

And there is no reason to (indeed, please let's not) borrow the ways of the old brain, its lusts and hungers, cravings and angers, feelings and fears, which can drive us along paths seen as harmful by the new brain. Harmful at least from the perspective that

Hawkins and I, and almost certainly you, value. For he is very clear that our enlightened values must, and do, diverge sharply from the primary and primitive value of our selfish genes—the raw imperative to reproduce at all costs. Without an old brain, in his view (which I suspect may be controversial), there is no reason to expect an AI to harbour malevolent feelings toward us. By the same token, and also perhaps controversially, he doesn't think switching off a conscious AI would be murder: Without an old brain, why would it feel fear or sadness? Why would it want to survive?

In the chapter "Genes Versus Knowledge," we are left in no doubt about the disparity between the goals of old brain (serving selfish genes) and of the new brain (knowledge). It is the glory of the human cerebral cortex that it—unique among all animals and unprecedented in all geological time—has the power to defy the dictates of the selfish genes. We can enjoy sex without procreation. We can devote our lives to philosophy, mathematics, poetry, astrophysics, music, geology, or the warmth of human love, in defiance of the old brain's genetic urging that these are a waste of time—time that "should" be spent fighting rivals and pursuing multiple sexual partners: "As I see it, we have a profound choice to make. It is a choice between favoring the old brain or favoring the new brain. More specifically, do we want our future to be driven by the processes that got us here, namely, natural selection, competition, and the drive of selfish genes? Or, do we want our future to be driven by intelligence and its desire to understand the world?"

I began by quoting T. H. Huxley's endearingly humble remark on closing Darwin's *Origin*. I'll end with just one of Jeff Hawkins's many fascinating ideas—he wraps it up in a mere couple of pages—which had me echoing Huxley. Feeling the need for a cosmic tombstone, something to let the galaxy know that we were once here and capable of announcing the fact, Hawkins notes that all civilisations are ephemeral. On the scale of universal time, the interval between a civilisation's invention of electromagnetic communication and its extinction is like the flash of a firefly. The chance of any one flash coinciding with another is unhappily small. What we need, then—why I called it a tombstone—is a message that says not "We are here" but "We were once here."

And the tombstone must have cosmic-scale duration: not only must it be visible from parsecs away, it must last for millions if not billions of years, so that it is still proclaiming its message when other flashes of intellect intercept it long after our extinction. Broadcasting prime numbers or the digits of π won't cut it. Not as a radio signal or a pulsed laser beam, anyway. They certainly proclaim biological intelligence, which is why they are the stock-in-trade of SETI (the search for extraterrestrial intelligence) and science fiction, but they are too brief, too in the present. So, what signal would last long enough and be detectable from a very great distance in any direction? This is where Hawkins provoked my inner Huxley.

It's beyond us today, but in the future, before our firefly flash is spent, we could put into orbit around the Sun a series of satellites "that block a bit of the Sun's light in a pattern that would not occur naturally. These orbiting Sun blockers would continue to orbit the Sun for millions of years, long after we are gone, and they could be detected from far away." Even if the spacing of these umbral satellites is not literally a series of prime numbers, the message could be made unmistakable: "Intelligent Life Woz 'Ere."

What I find rather pleasing—and I offer the vignette to Jeff Hawkins to thank him for the pleasure his brilliant book has given me—is that a cosmic message coded in the form of a pattern of intervals between spikes (or in his case anti-spikes, as his satellites dim the Sun) would be using the same kind of code as a neuron.

This is a book about how the brain works. It works the brain in a way that is nothing short of exhilarating.

PART 1

A New Understanding of the Brain

The cells in your head are reading these words. Think of how remarkable that is. Cells are simple. A single cell can't read, or think, or do much of anything. Yet, if we put enough cells together to make a brain, they not only read books, they write them. They design buildings, invent technologies, and decipher the mysteries of the universe. How a brain made of simple cells creates intelligence is a profoundly interesting question, and it remains a mystery.

Understanding how the brain works is considered one of humanity's grand challenges. The quest has spawned dozens of national and international initiatives, such as Europe's Human Brain Project and the International Brain Initiative. Tens of thousands of neuroscientists work in dozens of specialties, in practically every country in the world, trying to understand the brain. Although neuroscientists study the brains of different animals and ask varied questions, the ultimate goal of neuroscience is to learn how the human brain gives rise to human intelligence.

You might be surprised by my claim that the human brain remains a mystery. Every year, new brain-related discoveries are announced, new brain books are published, and researchers in related fields such as artificial intelligence claim their creations are approaching the intelligence of, say, a mouse or a cat. It would be easy to conclude from this that scientists have a pretty good idea of how the brain works. But if you ask neuroscientists, almost all of them would admit that we are still in the dark. We have learned a tremendous amount of knowledge and facts about the brain, but we have little understanding of how the whole thing works.

In 1979, Francis Crick, famous for his work on DNA, wrote an essay about the state of brain science, titled "Thinking About the Brain." He described the large quantity of facts that scientists had collected about the brain, yet, he concluded, "in spite of the steady accumulation of detailed knowledge, how the human brain works is still profoundly mysterious." He went on to say, "What is conspicuously lacking is a broad framework of ideas in which to interpret these results."

Crick observed that scientists had been collecting data on the brain for decades. They knew a great many facts. But no one had figured out how to assemble those facts into something meaningful. The brain was like a giant jigsaw puzzle with thousands of pieces. The puzzle pieces were sitting in front of us, but we could not make sense of them. No one knew what the solution was supposed to look like. According to Crick, the brain was a mystery not because we hadn't collected enough data, but because we didn't know how to arrange the pieces we already had. In the forty years since Crick wrote his essay there have been many significant discoveries about the brain, several of which I will talk about later, but overall his observation is still true. How intelligence arises from cells in your head is still a profound mystery. As more puzzle pieces are collected each year, it sometimes feels as if we are getting further from understanding the brain, not closer.

I read Crick's essay when I was young, and it inspired me. I felt that we could solve the mystery of the brain in my lifetime, and I have pursued that goal ever since. For the past fifteen years, I have led a research team in Silicon Valley that studies a part of the brain called the neocortex. The neocortex occupies about 70 percent of the volume of a human brain and it is responsible for everything we associate with intelligence, from our senses of vision, touch, and hearing, to language in all its forms, to abstract thinking such as mathematics and philosophy. The goal of our research is to understand how the neocortex works in sufficient detail that we can explain the biology of the brain and build intelligent machines that work on the same principles.

In early 2016 the progress of our research changed dramatically. We had a breakthrough in our understanding. We realized that we and other scientists had missed a key ingredient.

With this new insight, we saw how the pieces of the puzzle fit together. In other words, I believe we discovered the framework that Crick wrote about, a framework that not only explains the basics of how the neocortex works but also gives rise to a new way to think about intelligence. We do not yet have a complete theory of the brain—far from it. Scientific fields typically start with a theoretical framework and only later do the details get worked out. Perhaps the most famous example is Darwin's theory of evolution. Darwin proposed a bold new way of thinking about the origin of species, but the details, such as how genes and DNA work, would not be known until many years later.

To be intelligent, the brain has to learn a great many things about the world. I am not just referring to what we learn in school, but to basic things, such as what everyday objects look, sound, and feel like. We have to learn how objects behave, from how doors open and close to what the apps on our smartphones do when we touch the screen. We need to learn where everything is located in the world, from where you keep your personal possessions in your home to where the library and post office are in your town. And of course, we learn higher-level concepts, such as the meaning of "compassion" and "government." On top of all this, each of us learns the meaning of tens of thousands of words. Every one of us possesses a tremendous amount of knowledge about the world. Some of our basic skills are determined by our genes, such as how to eat or how to recoil from pain. But most of what we know about the world is learned.

Scientists say that the brain learns a model of the world. The word "model" implies that what we know is not just stored as a pile of facts but is organized in a way that reflects the structure of the world and everything it contains. For example, to know what a bicycle is, we don't remember a list of facts about bicycles. Instead, our brain creates a model of bicycles that includes the different parts, how the parts are arranged relative to each other, and how the different parts move and work together. To recognize something, we need to first learn what it looks and feels like, and to achieve goals we need to learn how things in the world typically behave when we interact with them. Intelligence is intimately tied to the brain's model of the world; therefore, to understand how the brain creates intelligence, we have to figure out how the brain,

made of simple cells, learns a model of the world and everything in it.

Our 2016 discovery explains how the brain learns this model. We deduced that the neocortex stores everything we know, all our knowledge, using something called reference frames. I will explain this more fully later, but for now, consider a paper map as an analogy. A map is a type of model: a map of a town is a model of the town, and the grid lines, such as lines of latitude and longitude, are a type of reference frame. A map's grid lines, its reference frame, provide the structure of the map. A reference frame tells you where things are located relative to each other, and it can tell you how to achieve goals, such as how to get from one location to another. We realized that the brain's model of the world is built using maplike reference frames. Not one reference frame, but hundreds of thousands of them. Indeed, we now understand that most of the cells in your neocortex are dedicated to creating and manipulating reference frames, which the brain uses to plan and think.

With this new insight, answers to some of neuroscience's biggest questions started to come into view. Questions such as, How do our varied sensory inputs get united into a singular experience? What is happening when we think? How can two people reach different beliefs from the same observations? And why do we have a sense of self?

This book tells the story of these discoveries and the implications they have for our future. Most of the material has been published in scientific journals. I provide links to these papers at the end of the book. However, scientific papers are not well suited for explaining large-scale theories, especially in a way that a nonspecialist can understand.

I have divided the book into three parts. In the first part, I describe our theory of reference frames, which we call the Thousand Brains Theory. The theory is partly based on logical deduction, so I will take you through the steps we took to reach our conclusions. I will also give you a bit of historical background to help you see how the theory relates to the history of thinking about the brain. By the end of the first part of the book, I hope you will have an understanding of what is going on in your head as you think and act within the world, and what it means to be intelligent.

The second part of the book is about machine intelligence. The twenty-first century will be transformed by intelligent machines in the same way that the twentieth century was transformed by computers. The Thousand Brains Theory explains why today's AI is not yet intelligent and what we need to do to make truly intelligent machines. I describe what intelligent machines in the future will look like and how we might use them. I explain why some machines will be conscious and what, if anything, we should do about it. Finally, many people are worried that intelligent machines are an existential risk, that we are about to create a technology that will destroy humanity. I disagree. Our discoveries illustrate why machine intelligence, on its own, is benign. But, as a powerful technology, the risk lies in the ways humans might use it.

In the third part of the book, I look at the human condition from the perspective of the brain and intelligence. The brain's model of the world includes a model of our self. This leads to the strange truth that what you and I perceive, moment to moment, is a simulation of the world, not the real world. One consequence of the Thousand Brains Theory is that our beliefs about the world can be false. I explain how this can occur, why false beliefs can be difficult to eliminate, and how false beliefs combined with our more primitive emotions are a threat to our long-term survival.

The final chapters discuss what I consider to be the most important choice we will face as a species. There are two ways to think about ourselves. One is as biological organisms, products of evolution and natural selection. From this point of view, humans are defined by our genes, and the purpose of life is to replicate them. But we are now emerging from our purely biological past. We have become an intelligent species. We are the first species on Earth to know the size and age of the universe. We are the first species to know how the Earth evolved and how we came to be. We are the first species to develop tools that allow us to explore the universe and learn its secrets. From this point of view, humans are defined by our intelligence and our knowledge, not by our genes. The choice we face as we think about the future is, should we continue to be driven by our biological past or choose instead to embrace our newly emerged intelligence?

We may not be able to do both. We are creating powerful technologies that can fundamentally alter our planet, manipulate

biology, and soon, create machines that are smarter than we are. But we still possess the primitive behaviors that got us to this point. This combination is the true existential risk that we must address. If we are willing to embrace intelligence and knowledge as what defines us, instead of our genes, then perhaps we can create a future that is longer lasting and has a more noble purpose.

The journey that led to the Thousand Brains Theory has been long and convoluted. I studied electrical engineering in college and had just started my first job at Intel when I read Francis Crick's essay. It had such a profound effect on me that I decided to switch careers and dedicate my life to studying the brain. After an unsuccessful attempt to get a position studying brains at Intel, I applied to be a graduate student at MIT's AI lab. (I felt that the best way to build intelligent machines was first to study the brain.) In my interviews with MIT faculty, my proposal to create intelligent machines based on brain theory was rejected. I was told that the brain was just a messy computer and there was no point in studying it. Crestfallen but undeterred, I next enrolled in a neuroscience PhD program at the University of California, Berkeley. I started my studies in January 1986.

Upon arriving at Berkeley, I reached out to the chair of the graduate group of neurobiology, Dr. Frank Werblin, for advice. He asked me to write a paper describing the research I wanted to do for my PhD thesis. In the paper, I explained that I wanted to work on a theory of the neocortex. I knew that I wanted to approach the problem by studying how the neocortex makes predictions. Professor Werblin had several faculty members read my paper, and it was well received. He told me that my ambitions were admirable, my approach was sound, and the problem I wanted to work on was one of the most important in science, but—and I didn't see this coming—he didn't see how I could pursue my dream at that time. As a neuroscience graduate student, I would have to work for a professor, doing similar work to what the professor was already working on. And no one at Berkeley, or anywhere else that he knew of, was doing something close enough to what I wanted to do.

Trying to develop an overall theory of brain function was considered too ambitious and therefore too risky. If a student worked on this for five years and didn't make progress, they might

not graduate. It was similarly risky for professors; they might not get tenure. The agencies that dispensed funding for research also thought it was too risky. Research proposals that focused on theory were routinely rejected.

I could have worked in an experimental lab, but after interviewing at a few I knew that it wasn't a good fit for me. I would be spending most of my hours training animals, building experimental equipment, and collecting data. Any theories I developed would be limited to the part of the brain studied in that lab.

For the next two years, I spent my days in the university's libraries reading neuroscience paper after neuroscience paper. I read hundreds of them, including all the most important papers published over the previous fifty years. I also read what psychologists, linguists, mathematicians, and philosophers thought about the brain and intelligence. I got a first-class, albeit unconventional, education. After two years of self-study, a change was needed. I came up with a plan. I would work again in industry for four years and then reassess my opportunities in academia. So, I went back to working on personal computers in Silicon Valley.

I started having success as an entrepreneur. From 1988 to 1992, I created one of the first tablet computers, the GridPad. Then in 1992, I founded Palm Computing, beginning a ten-year span when I designed some of the first handheld computers and smartphones such as the PalmPilot and the Treo. Everyone who worked with me at Palm knew that my heart was in neuroscience, that I viewed my work in mobile computing as temporary. Designing some of the first handheld computers and smartphones was exciting work. I knew that billions of people would ultimately rely on these devices, but I felt that understanding the brain was even more important. I believed that brain theory would have a bigger positive impact on the future of humanity than computing. Therefore, I needed to return to brain research.

There was no convenient time to leave, so I picked a date and walked away from the businesses I helped create. With the assistance and prodding of a few neuroscientist friends, (notably Bob Knight at UC Berkeley, Bruno Olshausen at UC Davis, and Steve Zornetzer at NASA Ames Research), I created the Redwood Neuroscience Institute (RNI) in 2002. RNI focused exclusively on

neocortical theory and had ten full-time scientists. We were all interested in large-scale theories of the brain, and RNI was one of the only places in the world where this focus was not only tolerated but expected. Over the course of the three years that I ran RNI, we had over one hundred visiting scholars, some of whom stayed for days or weeks. We had weekly lectures, open to the public, which usually turned into hours of discussion and debate.

Everyone who worked at RNI, including me, thought it was great. I got to know and spend time with many of the world's top neuroscientists. It allowed me to become knowledgeable in multiple fields of neuroscience, which is difficult to do with a typical academic position. The problem was that I wanted to know the answers to a set of specific questions, and I didn't see the team moving toward consensus on those questions. The individual scientists were content to do their own thing. So, after three years of running an institute, I decided the best way to achieve my goals was to lead my own research team.

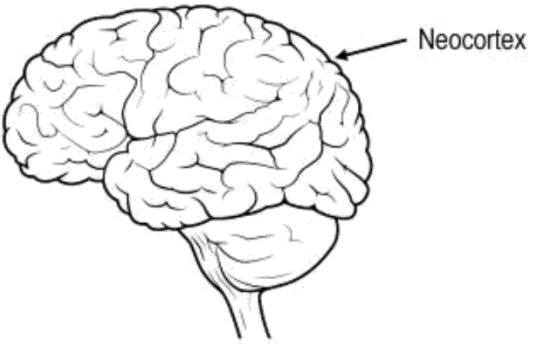
RNI was in all other ways doing well, so we decided to move it to UC Berkeley. Yes, the same place that told me I couldn't study brain theory decided, nineteen years later, that a brain-theory center was exactly what they needed. RNI continues today as the Redwood Center for Theoretical Neuroscience.

As RNI moved to UC Berkeley, several colleagues and I started Numenta. Numenta is an independent research company. Our primary goal is to develop a theory of how the neocortex works. Our secondary goal is to apply what we learn about brains to machine learning and machine intelligence. Numenta is similar to a typical research lab at a university, but with more flexibility. It allows me to direct a team, make sure we are all focused on the same task, and try new ideas as often as needed.

As I write, Numenta is over fifteen years old, yet in some ways we are still like a start-up. Trying to figure out how the neocortex works is extremely challenging. To make progress, we need the flexibility and focus of a start-up environment. We also need a lot of patience, which is not typical for a start-up. Our first significant discovery—how neurons make predictions—occurred in 2010, five years after we started. The discovery of maplike reference frames in the neocortex occurred six years later in 2016.

In 2019, we started to work on our second mission, applying

millimeters). It wraps around the older parts of the brain such that when you look at a human brain, most of what you see is the neocortex (with its characteristic folds and creases), with bits of the old brain and the spinal cord sticking out the bottom.



A human brain

The neocortex is the organ of intelligence. Almost all the capabilities we think of as intelligence—such as vision, language, music, math, science, and engineering—are created by the neocortex. When we think about something, it is mostly the neocortex doing the thinking. Your neocortex is reading or listening to this book, and my neocortex is writing this book. If we want to understand intelligence, then we have to understand what the neocortex does and how it does it.

An animal doesn't need a neocortex to live a complex life. A crocodile's brain is roughly equivalent to our brain, but without a proper neocortex. A crocodile has sophisticated behaviors, cares for its young, and knows how to navigate its environment. Most people would say a crocodile has some level of intelligence, but nothing close to human intelligence.

The neocortex and the older parts of the brain are connected via nerve fibers; therefore, we cannot think of them as completely separate organs. They are more like roommates, with separate agendas and personalities, but who need to cooperate to get anything done. The neocortex is in a decidedly unfair position, as it doesn't control behavior directly. Unlike other parts of the brain, none of the cells in the neocortex connect directly to muscles, so it can't, on its own, make any muscles move. When the neocortex wants to do something, it sends a signal to the old brain,

in a sense asking the old brain to do its bidding. For example, breathing is a function of the brain stem, requiring no thought or input from the neocortex. The neocortex can temporarily control breathing, as when you consciously decide to hold your breath. But if the brain stem detects that your body needs more oxygen, it will ignore the neocortex and take back control. Similarly, the neocortex might think, "Don't eat this piece of cake. It isn't healthy." But if older and more primitive parts of the brain say, "Looks good, smells good, eat it," the cake can be hard to resist. This battle between the old and new brain is an underlying theme of this book. It will play an important role when we discuss the existential risks facing humanity.

The old brain contains dozens of separate organs, each with a specific function. They are visually distinct, and their shapes, sizes, and connections reflect what they do. For example, there are several pea-size organs in the amygdala, an older part of the brain, that are responsible for different types of aggression, such as premeditated and impulsive aggression.

The neocortex is surprisingly different. Although it occupies almost three-quarters of the brain's volume and is responsible for a myriad of cognitive functions, it has no visually obvious divisions. The folds and creases are needed to fit the neocortex into the skull, similar to what you would see if you forced a napkin into a large wine glass. If you ignore the folds and creases, then the neocortex looks like one large sheet of cells, with no obvious divisions.

Nonetheless, the neocortex is still divided into several dozen areas, or regions, that perform different functions. Some of the regions are responsible for vision, some for hearing, and some for touch. There are regions responsible for language and planning. When the neocortex is damaged, the deficits that arise depend on what part of the neocortex is affected. Damage to the back of the head results in blindness, and damage to the left side could lead to loss of language.

The regions of the neocortex connect to each other via bundles of nerve fibers that travel under the neocortex, the so-called white matter of the brain. By carefully following these nerve fibers, scientists can determine how many regions there are and how they are connected. It is difficult to study human brains, so the first


image
not
available


image
not
available


image
not
available

that seem to have little order, suggesting that information goes all over at once. All regions, no matter what function they perform, look similar in detail to all other regions.

We will meet the first person who made sense of these observations in the next chapter.

This is a good point to say a few words about the style of writing in this book. I am writing for an intellectually curious lay reader. My goal is to convey everything you need to know to understand the new theory, but not a lot more. I assume most readers will have limited prior knowledge of neuroscience. However, if you have a background in neuroscience, you will know where I am omitting details and simplifying complex topics. If that applies to you, I ask for your understanding. There is an annotated reading list at the back of the book where I discuss where to find more details for those who are interested.