KNOW THYSELF

The Science of Self-Awareness

STEPHEN M. FLEMING

BASIC BOOKS
New York

CONTENTS

For Helen and Finn

BASIC
BOOKS

PREFACE

Imagine you arrive at your doctor's office for an appointment to discuss some recent chest pains. You undergo a series of blood tests and scans, and a week later you return to the clinic, where your doctor reviews the results with you. The condition seems serious, and she briskly recommends surgery for a heart bypass. When you ask your doctor why she is confident the procedure is necessary, she walks you through her thought process, including the possibility that she is wrong and what being wrong might entail, before reiterating her advice that you undergo the surgery. What would you do?

Now imagine that, after you undergo a series of blood tests and scans, the data is fed into an artificially intelligent assistant, which confidently states that the condition seems serious and it would be desirable if you had surgery for a heart bypass. When you ask your doctor whether this is really necessary, she can't tell you; she doesn't know why the recommendation has been made. All she can say is that, when fed the full range of test data, the AI has been highly accurate in the past, and that it would be wise to trust it and proceed with the surgery. What would you do?

In the first case, the answer probably seems obvious: if the doctor is confident and able to explain her reasons for being confident, you feel you should trust her advice. In the second, however, it may not be so clear. Many of us intuitively feel that if a person or a machine is going to be making high-stakes decisions on our behalf, we should be able to ask them to explain *why* they have come up with a particular answer. Many of our legal frameworks—those that ascribe liability and blame for errors—are based on the notion of being able to justify and defend what we did and why we did it. Without an explanation, we are left with blind trust—in each other, or in our machines. Ironically, some of the highest performing machine learning algorithms are often the least explainable. In contrast, humans are rapacious explainers of

what we are doing and why, a capacity that depends on our ability to reflect on, think about, and know things about ourselves, including how we remember, perceive, decide, think, and feel.

Psychologists have a special name for this kind of self-awareness: metacognition—literally, the ability to think about our own thinking, from the Greek "meta" meaning "after" or "beyond." Metacognition is a fragile, beautiful, and frankly bizarre feature of the human mind, one that has fascinated scientists and philosophers for centuries. In the biologist Carl Linnaeus's famous 1735 book *Systema Naturae*, he carefully noted down the physical features of hundreds of species. But when it came to our genus, *Homo*, he was so captivated with humans' ability for metacognition that he simply annotated his entry with the one-line Latin description "Nosce te ipsum"—those that know themselves.[1]

Self-awareness is a defining feature of human experience. Take a student, Jane, who is studying for an engineering exam. What might be going through her head? She is no doubt juggling a range of facts and formulas that she needs to master and understand. But she is also, perhaps without realizing it, figuring out how, when, and what to study. Which environment is better, a lively coffee shop or a quiet library? Does she learn best by rereading her notes or by practicing problem sets? Would it be better to shut the book on one topic and move onto another? Can she stop studying altogether and head out with friends?

Getting these decisions right is clearly critical for Jane's chances of success. She would not want to fall into a trap of thinking she knows a topic well when she does not, or to place her trust in a dodgy study strategy. But no one is giving her the answers to these questions. Instead, she is relying on her awareness of how she learns.

Our powers of self-reflection do not lose their importance when we leave the classroom or the exam hall. Consider the experience of James Nestor, an author and free diver. In his book *Deep*, Nestor recounts how he traveled to coastal locations in Greece and the Bahamas to report on free-diving tournaments. At each tournament, there is only one goal: to dive deeper than all the other competitors, all on a single breath. To prove that they have reached a particular depth, the divers retrieve a tag with a number emblazoned on it. If they pass out after surfacing, the dive

is declared null and void. To be successful, professional free divers must be acutely self-aware of their ability to reach a depth while avoiding injury or even death. Slight underconfidence will lead to underperformance, whereas slight overconfidence may be fatal. It's telling that a large part of free divers' training takes place on land, in psychological exploration of their underwater capacities and limitations.[2]

Or how about the case of Judith Keppel, one of the first contestants on the British TV game show *Who Wants to Be a Millionaire?* For each question, contestants are asked if they are sure they know the right answer and want to risk their existing winnings on the chance of a higher prize, or if they'd prefer to walk away with whatever they have already won. The stakes are high: being wrong means losing everything you have earned. In Keppel's case, she faced this decision with £500,000 on the line. The million-pound question was: "Which king was married to Eleanor of Aquitaine?" After a brief discussion with the show's host, Chris Tarrant, she settled on the answer of Henry II. Then Tarrant asked his killer question, the moment when contestants typically agonize the most: "Is that your final answer?" Success again rests on self-awareness. You want to know if you're likely to be right before accepting the gamble. Keppel stuck to her guns and became the show's first winner of the top prize.

What unites the stories of Jane, James, and Judith is how keenly their success or failure hinges on having accurate self-awareness. To appreciate the power of metacognition, we can reimagine their stories in a world where self-awareness is inaccurate. Jane might have erroneously thought that because the fluid mechanics problems felt easy, she could close the book on that topic and move on. She would think she was doing fine, even if this was not the case. A metacognitive error such as this could lead to catastrophic failure in the exam, despite Jane's raw ability and diligent studying. In Judith's case, we can identify two types of metacognitive failure: She may have known the answer but thought she did not, and therefore would have missed out on the opportunity to become a millionaire. Or she may have been overconfident, choosing to gamble on a wrong answer and losing everything. In James's case, such overconfidence may even be the difference between life and death. If he had thought that he was

able to handle deeper depths than he was capable of, he would, like a submarine Icarus, have overreached and realized his mistake only when it was too late.

We often overlook the power of metacognition in shaping our own lives, both for good and ill. The relevance of good self-awareness can seem less obvious than, say, the ability to solve equations, perform athletic feats, or remember historical facts. For most of us, our metacognition is like the conductor of an orchestra, occasionally intervening to nudge and guide the players in the right (or wrong) direction in ways that are often unnoticed or unappreciated at the time. If the conductor was absent, the orchestra would continue to play—just as Jane, James, and Judith would continue to plow on with studying, diving, and game-show answering even if their self-awareness was temporarily abolished. But a good conductor can make the difference between a routine rehearsal and a world-class performance—just as the subtle influence of metacognition can make the difference between success and failure, or life and death.

Another reason why the role of self-awareness is sometimes ignored is that it has historically proven difficult to measure, define, and study. But this is now changing. A new branch of neuroscience—metacognitive neuroscience—is pulling back the veil on how the human mind self-reflects. By combining innovative laboratory tests with the latest in brain imaging technology, we are now gaining an increasingly detailed picture of how self-awareness works, both as a cognitive process and as a biological one. As we will see, a science of metacognition can take us further than ever before in knowing ourselves.[3]

Creating a Science of Self-Awareness

I have been fascinated by the puzzle of self-awareness ever since I was a teenager, when I was drawn to books on the brain and mind. I remember glancing up from one of those books while lying by a pool during a summer vacation and daydreaming about my experience: Why should the mere activity of brain cells in my head lead to *this* unique experience of light shimmering on the surface of the swimming pool? And more to the point: How can the very

same brain that is having this experience allow me to think about these mysteries in the first place? It was one thing to be conscious, but to know I was conscious and to think about my own awareness —that's when my head began to spin. I was hooked.

I now run a neuroscience lab dedicated to the study of self-awareness at University College London. My team is one of several working within the Wellcome Centre for Human Neuroimaging, located in an elegant town house in Queen Square in London.[4] The basement of our building houses large machines for brain imaging, and each group in the Centre uses this technology to study how different aspects of the mind and brain work: how we see, hear, remember, speak, make decisions, and so on. The students and postdocs in my lab focus on the brain's capacity for self-awareness. I find it a remarkable fact that something unique about our biology has allowed the human brain to turn its thoughts on itself.

Until quite recently, however, this all seemed like nonsense. As the nineteenth-century French philosopher Auguste Comte put it: "The thinking individual cannot cut himself in two—one of the parts reasoning, while the other is looking on. Since in this case the organ observed and the observing organ are identical, how could any observation be made?"[5] In other words, how can the same brain turn its thoughts upon itself?

Comte's argument chimed with scientific thinking at the time. After the Enlightenment dawned on Europe, an increasingly popular view was that self-awareness was special and not something that could be studied using the tools of science. Western philosophers were instead using self-reflection as a philosophical tool, much as mathematicians use algebra in the pursuit of new mathematical truths. René Descartes relied on self-reflection in this way to reach his famous conclusion "I think, therefore I am," noting along the way that "I know clearly that there is nothing that can be perceived by me more easily or more clearly than my own mind." Descartes proposed that a central soul was the seat of thought and reason, commanding our bodies to act on our behalf. The soul could not be split in two—it just *was*. Self-awareness was therefore mysterious and indefinable, and off-limits to science.[6]

We now know that the premise of Comte's worry is false. The human brain is not a single, indivisible organ. Instead, the brain is

made up of billions of small components—neurons—that each crackle with electrical activity and participate in a wiring diagram of mind-boggling complexity. Out of the interactions among these cells, our entire mental life—our thoughts and feelings, hopes and dreams—flickers in and out of existence.

But rather than being a meaningless tangle of connections with no discernible structure, this wiring diagram also has a broader architecture that divides the brain into distinct regions, each engaged in specialized computations. Just as a map of a city need not include individual houses to be useful, we can obtain a rough overview of how different areas of the human brain are working together at the scale of regions rather than individual brain cells. Some areas of the cortex are closer to the inputs (such as the eyes) and others are further up the processing chain. For instance, some regions are primarily involved in seeing (the visual cortex, at the back of the brain), others in processing sounds (the auditory cortex), while others are involved in storing and retrieving memories (such as the hippocampus).

In a reply to Comte in 1865, the British philosopher John Stuart Mill anticipated the idea that self-awareness might also depend on the interaction of processes operating within a single brain and was thus a legitimate target of scientific study. Now, thanks to the advent of powerful brain imaging technologies such as functional magnetic resonance imaging (fMRI), we know that when we self-reflect, particular brain networks indeed crackle into life and that damage or disease to these same networks can lead to devastating impairments of self-awareness.[7]

Know Thyself Better

I often think that if we were not so thoroughly familiar with our own capacity for self-awareness, we would be gobsmacked that the brain is able to pull off this marvelous conjuring trick. Imagine for a moment that you are a scientist on a mission to study new life-forms found on a distant planet. Biologists back on Earth are clamoring to know what they're made of and what makes them tick. But no one suggests just asking them! And yet a Martian landing on Earth, after learning a bit of English or Spanish or

French, could do just that. The Martians might be stunned to find that we can already tell them something about what it is like to remember, dream, laugh, cry, or feel elated or regretful—all by virtue of being self-aware.[8]

But self-awareness did not just evolve to allow us to tell each other (and potential Martian visitors) about our thoughts and feelings. Instead, being self-aware is central to how we experience the world. We not only perceive our surroundings; we can also reflect on the beauty of a sunset, wonder whether our vision is blurred, and ask whether our senses are being fooled by illusions or magic tricks. We not only make decisions about whether to take a new job or whom to marry; we can also reflect on whether we made a good or bad choice. We not only recall childhood memories; we can also question whether these memories might be mistaken.

Self-awareness also enables us to understand that other people have minds like ours. Being self-aware allows me to ask, "How does this seem to me?" and, equally importantly, "How will this seem to someone else?" Literary novels would become meaningless if we lost the ability to think about the minds of others and compare their experiences to our own. Without self-awareness, there would be no organized education. We would not know who needs to learn or whether we have the capacity to teach them. The writer Vladimir Nabokov elegantly captured this idea that self-awareness is a catalyst for human flourishing:

> Being aware of being aware of being. In other words, if I not only know that I *am* but also know that I know it, then I belong to the human species. All the rest follows—the glory of thought, poetry, a vision of the universe. In that respect, the gap between ape and man is immeasurably greater than the one between amoeba and ape.[9]

In light of these myriad benefits, it's not surprising that cultivating accurate self-awareness has long been considered a wise and noble goal. In Plato's dialogue *Charmides*, Socrates has just returned from fighting in the Peloponnesian War. On his way home, he asks a local boy, Charmides, if he has worked out the

meaning of *sophrosyne*—the Greek word for temperance or moderation, and the essence of a life well lived. After a long debate, the boy's cousin Critias suggests that the key to *sophrosyne* is simple: self-awareness. Socrates sums up his argument: "Then the wise or temperate man, and he only, will know himself, and be able to examine what he knows or does not know.... No other person will be able to do this."[10]

Likewise, the ancient Greeks were urged to "know thyself" by a prominent inscription carved into the stone of the Temple of Delphi. For them, self-awareness was a work in progress and something to be striven toward. This view persisted into medieval religious traditions: for instance, the Italian priest and philosopher Saint Thomas Aquinas suggested that while God knows Himself by default, we need to put in time and effort to know our own minds. Aquinas and his monks spent long hours engaged in silent contemplation. They believed that only by participating in concerted self-reflection could they ascend toward the image of God.[11]

A similar notion of striving toward self-awareness is seen in Eastern traditions such as Buddhism. The spiritual goal of enlightenment is to dissolve the ego, allowing more transparent and direct knowledge of our minds acting in the here and now. The founder of Chinese Taoism, Lao Tzu, captured this idea that gaining self-awareness is one of the highest pursuits when he wrote, "To know that one does not know is best; Not to know but to believe that one knows is a disease."[12]

Today, there is a plethora of websites, blogs, and self-help books that encourage us to "find ourselves" and become more self-aware. The sentiment is well meant. But while we are often urged to have better self-awareness, little attention is paid to how self-awareness actually works. I find this odd. It would be strange to encourage people to fix their cars without knowing how the engine worked, or to go to the gym without knowing which muscles to exercise. This book aims to fill this gap. I don't pretend to give pithy advice or quotes to put on a poster. Instead, I aim to provide a guide to the building blocks of self-awareness, drawing on the latest research from psychology, computer science, and neuroscience. By understanding how self-awareness works, I aim to put us in a position to answer the Athenian call to use it better.

I also aim to help us use our machines better—both those that exist today and those that are likely to arrive in the near future. As with your imagined visit to the doctor's artificially intelligent clinic and its inexplicable advice to have surgery, we are already being forced to deal with complex systems making decisions we do not understand. We are surrounded by smart but unconscious algorithms—from climate forecasting models to automatic financial traders—and similar tools are poised to encroach on all areas of our lives. In many cases, these algorithms make our lives easier and more productive, and they may be required to help us tackle unprecedented challenges such as climate change. But there is also a danger that deferring to supersmart black boxes will limit human autonomy: by removing metacognition from the equation, we will not know why or how certain decisions were made and instead be forced into blindly following the algorithms' advice. As the philosopher Daniel Dennett points out: "The real danger, I think, is not that machines more intelligent than we are will usurp our role as captains of our destinies, but that we will overestimate the comprehension of our latest thinking tools, prematurely ceding authority to them far beyond their competence."[13] As we will see, the science of self-awareness provides us with alternative visions of this future, ones that ensure that awareness of competence remains at the top of the priority list, both for ourselves and our machines.

Looking Ahead

Let's take a look at the road ahead. The big idea of this book is that the human brain plays host to specific algorithms for self-awareness. How these algorithms work will occupy us in Part I. We will see that the neural circuits supporting metacognition did not just pop up out of nowhere. Instead, they are grounded in the functions of the evolved human brain. This means that many of the building blocks of metacognition are also shared with other species and are in place early in human development. We'll cover both the unconscious processes that form the building blocks of self-monitoring and the conscious processes that enable you to be self-aware of the experiences you are having. As will become clear,

when we talk about self-awareness, what we really mean is a collection of capacities—such as being able to recognize our mistakes and comment on our experience—that when bundled together result in a self-aware human being.[14]

By the end of Part I, we will have seen how a number of critical components come together to create a fully-fledged capacity for self-awareness. We will also be in a position to understand how and why these processes sometimes go wrong, leading to failures of self-awareness in diseases such as schizophrenia and dementia. In Part II, we will then turn to how we use self-awareness in many areas of our lives to learn, make decisions, and collaborate with others. By understanding how and why self-awareness may become distorted—and by recognizing both its power and fragility —we will be in a position to ensure that we do not end up in situations in which it tends to misfire. We'll dig into several important arenas of human affairs—including the crucial role that metacognition plays in witnesses testimony, in politics, and in science—to see why knowing ourselves, and knowing how others know themselves, is crucial to building a fairer and better society. We'll explore how self-awareness helps us separate reality from imagination and how, by learning to harness it, it can even help us shape our dreams. We will see that because self-awareness is sometimes absent there are, in fact, plenty of cases in which we humans are also no better than black boxes, unable to explain what we have done or why.

We will also see that, despite these limitations, the human capacity for self-awareness and self-explanation is what underpins our notions of autonomy and responsibility. We'll explore the role of self-awareness in classroom learning and teaching. We'll see why in sports it might be better to be less self-aware to perform well but more self-aware when coaching others. We'll see how digital technology changes our awareness of ourselves and others in a range of crucial ways. Indeed, I'll make the case that in a world of increasing political polarization and misinformation, cultivating the ability to self-reflect and question our beliefs and opinions has never been more essential. We'll explore why computers—even the most powerful—currently lack metacognition, and how the increasing prevalence of machine learning in AI means that algorithms for intelligence are rapidly diverging from algorithms

for self-awareness. We'll examine what this might mean for society and how we might fix it, either by attempting to build self-awareness into our computers or by ensuring we can understand and use the machines that we build. However that effort concludes, it may hold the key to solving some of the most pressing problems in society.

By the end of all this, I hope it will be clear why, from ancient Athens to the boardroom of Amazon.com, cultivating self-awareness has always been essential to flourishing and success. But we are getting ahead of ourselves. To unravel the mysteries of how self-awareness works, we need to start with the simplest of building blocks. Let's begin with two features of how our minds work: how we track uncertainty and how we monitor our actions. These two features may appear simple, but they are fundamental components of a self-aware brain.

PART I

BUILDING MINDS THAT KNOW THEMSELVES

1

HOW TO BE UNCERTAIN

> The other fountain [of] ideas, is the perception of the operation of our own minds within us.... And though it be not sense, as having nothing to do with external objects, yet it is very like it, and might properly enough be called internal sense.
>
> **—JOHN LOCKE,**
> *Essay Concerning Human Understanding, Book II*

Is something there, or not? This was the decision facing Stanislav Petrov one early morning in September 1983. Petrov was a lieutenant colonel in the Soviet Air Defense Forces and in charge of monitoring early warning satellites. It was the height of the Cold War between the United States and Russia, and there was a very real threat that long-range nuclear missiles could be launched by either side. That fateful morning, the alarms went off in Petrov's command center, alerting him that five US missiles were on their way to Russia. Under the doctrine of mutually assured destruction, his job was to immediately report the attack to his superiors so they could launch a counterattack. Time was of the essence—within twenty-five minutes, the missiles would detonate on Soviet soil.[1]

But Petrov decided that the alert was unlikely to be a real missile. Instead, he called in a system malfunction. To him, it seemed more probable that the satellite was unreliable—that the blip on the radar screen was noise, not signal—than that the United States had sent over missiles in a surprise attack that would surely launch a nuclear war. After a nervous wait of several minutes, he was proved right. The false alarm had been triggered

by the satellites mistaking the sun's reflection off the tops of clouds for missiles scudding through the upper atmosphere.

Petrov saw the world in shades of gray and was willing to entertain uncertainty about what the systems and his senses were telling him. His willingness to embrace ambiguity and question what he was being told arguably saved the world from disaster. In this chapter, we will see that representing uncertainty is a key ingredient in our recipe for creating self-aware systems. The human brain is in fact an exquisite uncertainty-tracking machine, and the role of uncertainty in how brains work goes much deeper than the kind of high-stakes decision facing Petrov. Without an ability to estimate uncertainty, it is unlikely that we would be able to perceive the world at all—and a wonderful side benefit is that we can also harness it to doubt ourselves.

Inverse Problems and How to Solve Them

The reason Petrov's decision was difficult was that he had to separate out signal from noise. The same blip on the radar screen could be due to an actual missile or noise in the system. It is impossible to work out which from the characteristics of the blip alone. This is known as an inverse problem—so called because solving it requires inverting the causal chain and making a best guess about what is causing the data we are receiving. In the same way, our brains are constantly solving inverse problems, unsure about what is really out there in the world.

The reason for this is that the brain is locked inside a dark skull and has only limited contact with the outside world through the lo-fi information provided by the senses. Take the seemingly simple task of deciding whether a light was just flashed in a darkened room. If the light flash is made dim enough, then sometimes you will say the light is present even when it is absent. Because your eye and brain form a noisy system, the firing of neurons in your visual cortex is not exactly the same for each repetition of the stimulus. Sometimes, even when the light isn't flashed, random noise in the system will lead to high firing rates, just like the blip on Petrov's radar screen was caused by atmospheric noise. Because the brain doesn't know whether these high firing rates are caused by signal or noise, if your visual

cortical neurons are firing vigorously it will seem as though a light was flashed even if it wasn't.[2]

Because our senses—touch, smell, taste, sight, and hearing—each have access to only a small, noisy slice of reality, they must pool their resources to come up with a best guess about what is really out there. They are rather like the blind men in the ancient Indian parable. The one holding the elephant's leg says it must be a pillar; the one who feels the tail says it is like a rope; the one who feels the trunk says it is like a tree branch; the one who feels the ear says it is like a hand fan; the one who feels its belly says it is like a wall; and the one who feels the tusk says it is like a solid pipe. Eventually, a stranger wanders past and informs them that they are, in fact, all correct and the elephant has all the features they observed. They would do better to combine their perspectives, he says, rather than argue.

A mathematical framework known as Bayes's theorem provides a powerful tool for thinking about these kinds of problems. To see how Bayes's rule helps us solve inverse problems, we can play the following game. I have three dice, two of which are regular dice with the numbers 1 to 6 on their faces, and one of which is a trick die with either a 0 or 3 on every face. From behind a curtain, I'm going to roll all three dice at once and tell you the combined total. On each roll, I might choose to use a trick die that shows all 0s, or a trick die that shows all 3s. For instance, on my first roll I might roll 2, 4, and 0 (on the third, trick die) for a combined total of 6. Your task is to tell me your best guess about the identity of the trick die—either a 3 or a 0—based only on your knowledge of the total.[3]

In this game, the 0 or the 3 on the trick die stand in for the "hidden" states of the world: whether the missile was present in Petrov's dilemma, or whether the light was flashed in the case of our visual cortex neuron. Somehow, we need to go back from the noisy evidence we have received—the sum total of the three dice—and use this to work out the hidden state.

Sometimes this is easy. If I tell you the combined total is 4 or less, then you know that the third die must have been showing 0 to produce such a low sum. If the combined total is greater than 12 (two 6s plus a quantity more than 0), then you know for sure that the third die must have been showing 3. But what about quantities

between these extremes? What about a total of 6 or 8? This is trickier.

One way we might go about solving this game is by trial and error. We could roll the three dice ourselves many times, record the total, and observe the true state of the world: what was actually showing on the face of the third die on each roll.

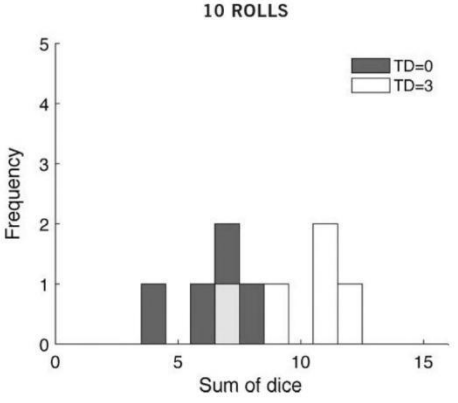
The first few rolls of the game might look like this:

Roll	Die 1	Die 2	Trick Die	Total
1	2	4	0	6
2	5	1	3	9
3	5	6	3	14

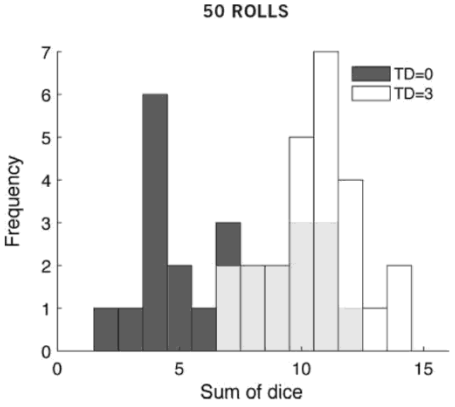
*

And so on, for many tens of rolls. An easier way to present this data is in a chart of the number of times we observe a particular total—say, 6—and the identity of the trick die at the time (0 or 3). We can select particular colors for the trick die number; here I've chosen gray for 0 and white for 3.

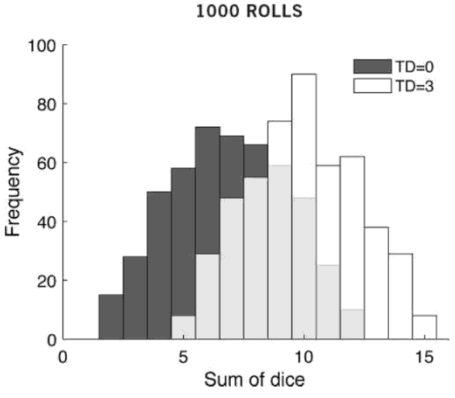
After ten rolls the graph might look like this:



This isn't very informative, and shows only a scatter of different totals, just like in our table. But after fifty rolls a pattern starts to emerge:



And after one thousand rolls, the pattern is very clear:



The counts from our experiment form two clear hills, with the majority falling in a middle range and peaks around 7 and 10. This makes sense. On average, the two real dice will give a total of around 7, and therefore adding either 0 or 3 from the trick die to this total will tend to give 7 or 10. And we see clear evidence for our intuition at the start: you only observe counts of 4 or less when the trick die equals 0, and you only observe counts of 13 or more when the trick die equals 3.

Now, armed with this data, let's return to our game. If I were to give you a particular total, such as 10, and ask you to guess what the trick die is showing, what should you answer? The graph above tells us that it is more likely that the total 10 is associated with the trick die having 3 on its face. From Bayes's rule, we know that the relative height of the white and gray bars (assuming we've performed our experiment a sufficient number of times) tells us precisely how much more likely the 3 is compared to the 0—in this

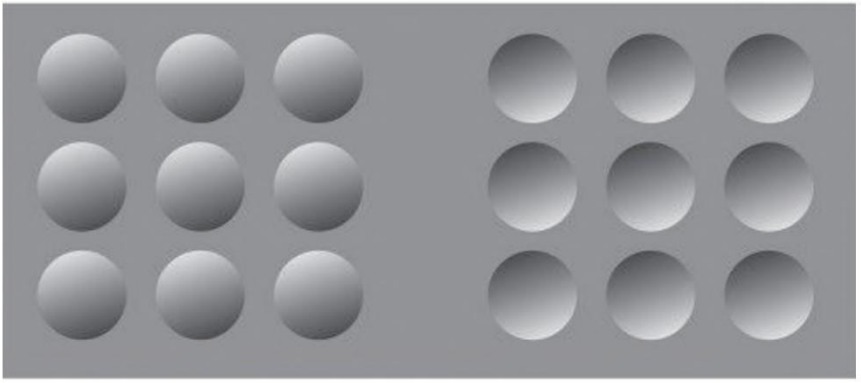
case, around twice as likely. The Bayes-optimal solution to this game is to always report the more likely value of the trick die, which amounts to saying 3 when the total is 9 or above and 0 when the total is 8 or lower.

What we've just sketched is an algorithm for making a decision from noisy information. The trick die is always lurking in the background because it is contributing to the total each time. But its true status is obscured by the noise added by the two normal dice, just as the presence of a missile could not be estimated by Petrov from the noisy radar signal alone. Our game is an example of a general class of problems involving decisions under uncertainty that can be solved by applying Bayes's rule.

In the case of Petrov's fateful decision, the set of potential explanations is limited: either there is a missile or it's a false alarm. Similarly, in our dice game, there are only two explanations to choose between: either the trick die is a 0 or it's a 3. But in most situations, not only is the sensory input noisy, but there is a range of potential explanations for the data streaming in through our senses. Imagine a drawing of a circle around twenty centimeters across and held at a distance of one meter from the eye. Light reflected from the circle travels in straight lines, passing through the lens of the eye and creating a small image (of a circle) on the retina. Because the image on the retina is two-dimensional, the brain could interpret it as being caused by any infinite number of circles of different sizes arranged at appropriate distances. Roughly the same retinal image would be caused by a forty-centimeter circle held at two meters, or an eight-meter circle at forty meters. In many cases, there is simply not enough information in the input to constrain what we see.

These more complex inverse problems can be solved by making guesses as to the best explanation based on additional information drawn from other sources. To estimate the actual diameter of the circle, for instance, we can use other cues such as differences in the images received by the two eyes, changes in the texture, position, and shading of nearby objects, and so on.

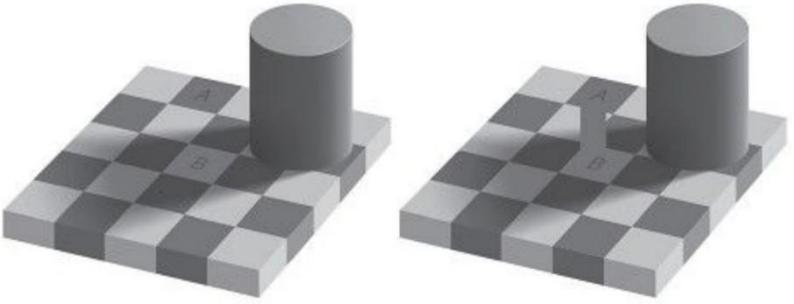
To experience this process in real time, take a look at these two pictures:



Most people see the image on the left-hand side as a series of bumps, raised above the surface. The image on the right, in contrast, looks like a series of little pits or depressions in the page. Why the difference?

The illusion is generated by your brain's solution to the inverse problem. The left and right sets of dots are actually the same image rotated 180 degrees (you can rotate the book to check!). The reason they appear different is that our visual system expects light to fall from above, because scenes are typically lit from sunlight falling from above our heads. In contrast, uplighting —such as when light from a fire illuminates the side of a cliff, or spotlights are projected upward onto a cathedral—is statistically less common. When viewing the two sets of dots, our brain interprets the lighter parts of the image on the left as being consistent with light striking a series of bumps and the darker parts of the image on the right as consistent with a series of shadows cast by holes, despite both being created from the same raw materials.

Another striking illusion is this image created by the vision scientist Edward Adelson:



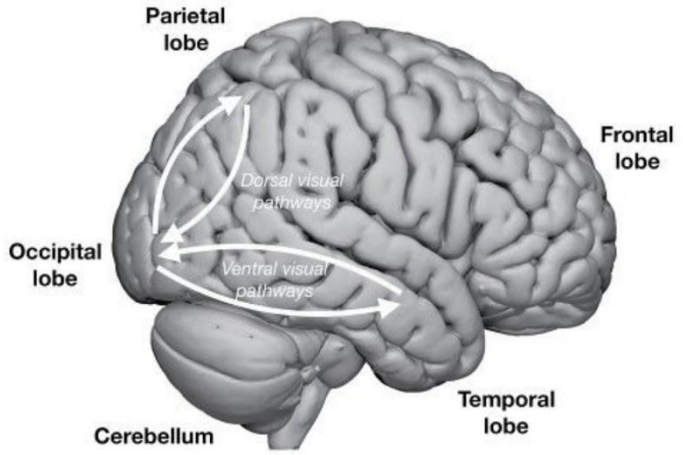
Adelson's checkerboard *(Edward H. Adelson)*

In the left-hand image, the squares labeled A and B are in fact identical shades of gray; they have the same luminance. Square B appears lighter because your brain "knows" that it is in shadow: in order to reflect the same level of light to the eye as A, which is fully illuminated, it must have started out lighter. The equivalence of A and B can be easily appreciated by connecting them up, as in the right-hand image. Now the cue provided by this artificial bridge overrides the influence of the shadow in the brain's interpretation of the squares (to convince yourself that the left- and right-hand images are the same, try using a sheet of paper to cover up the bottom half of the two images).

The upshot is that these surprising illusions are not really illusions at all. One interpretation of the image is given by our scientific instruments—the numbers produced by light meters and computer monitors. The other is provided by our visual systems that have been tuned to discover regularities such as shadows or light falling from above—regularities that help them build useful models of the world. In the real world, with light and shade and shadows, these models would usually be right. Many visual illusions are clever ways of exposing the workings of a system finely tuned for perceptual inference. And, as we will see in the next section, several principles of brain organization are consistent with this system solving inverse problems on a massive scale.

Building Models of the World

One of the best-understood parts of the human and monkey brain is the visual system. Distinct regions toward the back of the brain process different aspects of visual input, and each is labeled with increasing numbers for more advanced stages of image processing. V1 and V2 extract information about the orientation of lines and shapes, V4 about color, and V5 about whether objects are moving. Downstream of these V regions we hit regions of the ventral visual stream that are tasked with putting all these pieces together to identify whole objects, such as faces and bodies and tables and chairs. In parallel, the dorsal visual stream contains regions that specialize in keeping track of where things are and whether they are moving from place to place.[4]



The right hemisphere of the human brain. The locations of the four cortical lobes, the cerebellum, and key visual pathways are indicated.

At the start of the ventral visual stream, individual brain cells encode only a small amount of the external world, such as a patch in the lower left of our field of view. But as we move up the hierarchy, the cells begin to widen their focus, similar to a camera zooming out. By the time we reach the top of the hierarchy, where a stimulus is displayed matters much less than what it depicts—a face, house, cat, dog, etc. The lens is completely zoomed out, and information about the object's identity is represented independently of spatial location.

Crucially, however, information in the visual system does not just flow in one direction. For a long time, the dominant view of information processing in the brain was that it was a feed-forward system—taking in information from the outside world, processing it in hidden, complex ways, and then spitting out commands to make us walk and talk. This model has now been superseded by a raft of evidence that is difficult to square with the input-output view. In the visual system, for instance, there are just as many, if not more, connections in the reverse direction, known as feedback or top-down connections. Information travels both forward and backward; upper levels of the hierarchy both receive inputs from lower levels and send information back down in constant loops of neural activity. This way of thinking about the mind is known as predictive processing, and it represents a radically different understanding of what the brain does—although one with a long intellectual history, as the range of references in the endnote makes clear.[5]

Predictive processing architectures are particularly well suited to solving inverse problems. Instead of just passively taking in information, the brain can harness these top-down connections to actively construct our perception of the external world and shape what we see, hear, think, and feel. Higher levels furnish information about the type of things we might encounter in any given situation and the range of hypotheses we might entertain. For instance, you might know that your friend owns a Labrador, and so you expect to see a dog when you walk into the house but don't know exactly where in your visual field the dog will appear. This higher-level prior—the spatially invariant concept of "dog"—provides the relevant context for lower levels of the visual system to easily interpret dog-shaped blurs that rush toward you as you open the door.

The extent to which our perceptual systems should rely on these regularities—known as priors—is in turn dependent on how uncertain we are about the information being provided by our senses. Think back to Petrov's dilemma. If he was sure that his missile-detection technology was flawless and never subject to error, he would have been less willing to question what the system was telling him. Whether we should adjust our beliefs upon receiving new data depends on how reliable we think that

information is.

In fact, Bayesian versions of predictive processing tell us that we should combine different sources of information—our prior beliefs and the data coming in through our senses—in inverse proportion to how uncertain we are about them. We can think of this process as being similar to pouring cake batter into a flexible mold. The shape of the mold represents our prior assumptions about the world. The batter represents the sensory data—the light and sound waves hitting the eyes and ears. If the incoming data is very precise or informative, then the batter is very thick, or almost solid, and will be hardly affected by the shape of the mold (the priors). If, in contrast, the data is less precise, then the batter will be runnier, and the shape of the mold will dominate the shape of the final product.

For instance, our eyes provide more precise information about the location of objects than our hearing. This means that vision can act as a useful constraint on the location of a sound source, biasing our perception of where the sound is coming from. This is used to great effect by ventriloquists, who are seemingly able to throw their voices to a puppet held at arm's length. The real skill of ventriloquism is the ability to speak without moving the mouth. Once this is achieved, the brains of the audience do the rest, pulling the sound to its next most likely source, the talking puppet.[6]

It makes sense, then, that keeping track of uncertainty is an inherent part of how the brain processes sensory information. Recordings of cells from the visual cortex show us how this might be done. It's well known that moving objects such as a waving hand or a bouncing ball will activate neurons in an area of the monkey brain known as MT (the human equivalent is V5). But cells in MT do not just activate for any direction of movement. Some cells fire most strongly for objects moving to the left, others for up, down, and all other points of the compass. When firing rates of MT cells are recorded over multiple presentations of different motion directions, they begin to form a distribution like the ones we saw in our dice game. At any moment in time, these populations of MT cells can be thought of as signaling the uncertainty about a particular direction of motion, just as our noisy dice total signaled the probability of the trick die being a 0 or

a 3.[7]

Uncertainty is also critical for estimating the states of our own bodies. Information about where our limbs are in space, how fast our heart is beating, or the intensity of a painful stimulus is conveyed to the skull by sensory neurons. From the brain's perspective, there is little difference between the electrical impulses traveling down the optic nerve and the neural signals ascending from our gut, heart, muscles, or joints. They are all signals of what might be happening outside of the skull, and these signals are subject to illusions of the kind that we encountered for vision. In one famous experiment, stroking a rubber hand in time with the participant's own (hidden) hand is sufficient to convince the participant that the rubber hand is now their own. In turn, the illusion of ownership of the new rubber hand leads the brain to wind down the neural signals being sent to the actual hand. Just as the voice is captured by the ventriloquist's dummy, the synchrony with which the rubber hand is seen and felt to be stroked pulls the sense of ownership away from the real hand.[8]

Harnessing Uncertainty to Doubt Ourselves

Of course, no one is suggesting that we consciously churn through Bayesian equations every time we perceive the world. Instead, the machinery our brains use to solve inverse problems is applied without conscious thought, in what the German physicist Hermann von Helmholtz called a process of "unconscious inference." Our brains rapidly estimate the effects of light and shade on the dips, bumps, and checkerboards we encountered in the images on previous pages, literally in the blink of an eye. In a similar fashion, we reconstruct the face of a close friend, the taste of a fine wine, and the smell of freshly baked bread by combining priors and data, carefully weighting them by their respective uncertainties. Our perception of the world is what the neuroscientist Anil Seth refers to as a "controlled hallucination"—a best guess of what is really out there.

It is clear that estimating uncertainty about various sources of information is fundamental to how we perceive the world. But there is a remarkable side benefit of these ingenious solutions to

the inverse problem. In estimating uncertainty in order to perceive the world, we also gain the ability to doubt what we perceive. To see how easy it is to turn uncertainty into self-doubt, let's consider the dice game again. As the numbers in the game tend toward either 15 or 0, we become surer about the trick die showing a 3 or 0, respectively. But in the middle part of the graph, where the gray and white bars are of similar height—totals of 7 or 8—there is limited support for either option. If I ask you how confident you are about your response, it would be sensible to doubt decisions about the numbers 7 and 8 and to be more confident about smaller and larger quantities. In other words, we know that we are likely to know the answer when uncertainty is low, and we know that we are likely to *not* know the answer when uncertainty is high.

Bayes's rule provides us with a mathematical framework for thinking about these estimates of uncertainty, sometimes known as type 2 decisions—so called because they are decisions about the accuracy of other decisions, rather than type 1 decisions, which are about things in the world. Bayes's theorem tells us that it is appropriate to be more uncertain about responses toward the middle of the graph, because they are the ones most likely to result in errors and are associated with the smallest probability of being correct. Conversely, as we go out toward the tails of each distribution, the probability of being correct goes up. By harnessing the uncertainty that is inherent to solving inverse problems, we gain a rudimentary form of metacognition for free—no additional machinery is needed.[9]

And, because tracking uncertainty is foundational to how brains perceive the world, it is not surprising that this form of metacognition is widespread among a range of animal species. One of the first—and most ingenious—experiments on animal metacognition was developed by the psychologist J. David Smith in his study of a bottlenose dolphin named Natua. Smith trained Natua to press two different levers in his tank to indicate whether a sound was high-pitched or low-pitched. The low-pitched sound varied in frequency from very low to relatively high, almost as high as the high-pitched sound. There was thus a zone of uncertainty in which it wasn't always clear whether low or high was the right answer, just like in our dice game.[10]

Once Natua had got the hang of this task, a third lever was introduced into the tank that could be pressed to skip the current trial and move on to the next one—the dolphin equivalent of skipping a question on a multiple-choice quiz. Smith reasoned that if Natua declined to take on decisions when his uncertainty about the answer was high, he would be able to achieve a higher accuracy overall than if he was forced to guess. And this is exactly what Smith found. The data showed that Natua pressed the third lever mostly when the sound was ambiguous. As Smith recounts, "When uncertain, the dolphin clearly hesitated and wavered between his two possible responses, but when certain, he swam towards his chosen response so fast that his bow wave would soak the researchers' electronic switches."[11]

Macaque monkeys—which are found across Asia (and are fond of stealing tourists' food at temples and shrines)—also easily learn to track their uncertainty in a similar setup. In one experiment, macaques were trained to judge which was the biggest shape on a computer screen, followed by another choice between two icons. One icon led to a risky bet (three food pellets if right, or the removal of food if wrong), while the other, safe option always provided one food pellet—the monkey version of *Who Wants to Be a Millionaire?* The monkeys selected the risky option more often when they were correct, a telltale sign of metacognition. Even more impressively, they were able to immediately transfer these confidence responses to a new memory test without further training, ruling out the idea that they had just learned to associate particular stimuli with different confidence responses. Adam Kepecs's lab, based at Cold Spring Harbor in New York, has used a version of this task to show that rats also have a sense of whether they are likely to be right or wrong about which of two perfumes is most prominent in a mixture of odors. There is even some evidence to suggest that birds can transfer their metacognitive competence between different tests, just like monkeys.[12]

If a sensitivity to uncertainty is a fundamental property of how brains work, it makes sense that this first building block of metacognition might also be found early in the lives of human babies. Taking inspiration from Smith's tests, Louise Goupil and Sid Kouider at the École Normale Supérieure in Paris set out to measure how eighteen-month-old infants track uncertainty about

their decisions. While sitting on their mothers' laps, the babies were shown an attractive toy and allowed to play with it to whet their appetite for more playtime in the future. They then saw the toy being hidden in one of two boxes. Finally, after a brief delay, they were allowed to reach inside either of the boxes to retrieve the toy.

In reality, the toy was sneakily removed from the box by the experimenter. This allowed the researchers to measure the infants' confidence about their choice of box. They reasoned that, if the babies knew whether they were making a good or bad choice, they would be more willing to search for the (actually nonexistent) toy when the correct box was chosen compared to when they chose incorrectly. This was indeed the case: when babies made wrong moves, they were less persistent in searching for the toy. They were also more likely to ask their mother for help in retrieving the toy when they were most prone to making an error. This data tells us that even at a young age, infants can estimate how uncertain they are about simple choices, asking for help only when they most need it.[13]

We cannot know for sure how animals and babies are solving these problems, because—unlike human adults—they cannot tell us about what they are thinking and feeling. A critic could argue that they are following a lower-level rule that is shared across all the tasks in the experiment—something like, If I take a long time to decide, then I should press the "uncertain" key—without forming any feeling of uncertainty about the decisions they are making. In response to this critique, ever more ingenious experiments have been designed to rule out a variety of non-metacognitive explanations. For instance, to rule out tracking response time, other studies have given the animals the chance to bet on their choices before they have started the test and before response-time cues are available. In this setup, macaque monkeys are more likely to be correct when they choose to take the test than when they decline, suggesting that they know when they know the answer—a hallmark of metacognition.[14]

There is also evidence that otherwise intelligent species fail to track uncertainty in these situations, suggesting that feelings of uncertainty might really be picking up on the first glimmers of self-awareness, rather than a more generic cognitive ability.

Capuchin monkeys, a New World species found in South America, share many characteristics with macaques, using tools such as stones to crack open palm nuts, and living in large social groups. But capuchins appear unable to signal that they are uncertain in Smith's task. In a clever twist, it is possible to show that capuchins have no difficulty using a third response key to classify a new stimulus, but they are unable to use the same response to indicate when they are uncertain. This data suggests that when comparing two similar species of monkey, one may show signs of metacognition while another may not.[15]

Once uncertainty tracking is in place, it opens the door to a range of useful behaviors. For starters, being able to estimate uncertainty means we can use it to decide whether or not we need more information. Let's go back to our dice game. If I were to give you a total near the middle of the graph—a 7 or an 8—then you might reasonably be uncertain about whether to answer 0 or 3. Instead, you might ask me to roll the dice again. If I were to then roll a 5, a 4, and a 7, all with the same three dice, then you would be much more confident that the trick die was a 0. As long as each roll is independent of the previous one, Bayes's theorem tells us we can compute the probability that the answer is a 3 or a 0 by summing up the logarithm of the ratio of our confidence in each hypothesis after each individual roll.[16]

The brilliant British mathematician Alan Turing used this trick to figure out whether or not to change tack while trying to crack the German Enigma code in the Second World War. Each morning, his team would try new settings of their Enigma machine in an attempt to decode intercepted messages. The problem was how long to keep trying a particular pair of ciphers before discarding it and trying another. Turing showed that by accumulating multiple samples of information over time, the code breakers could increase their confidence in a particular setting being correct—and, critically, minimize the amount of time wasted testing the wrong ciphers.[17]

In the same way, we can use our current estimate of confidence to figure out whether a new piece of information will be helpful. If I get a 12 on my first roll, then I can be reasonably confident that the trick die is showing a 3 and don't need to ask for the dice to be rolled again. But if I get a 7 or 8, then it would be

prudent to roll again and resolve my current uncertainty about the right answer. The role of confidence in guiding people's decisions to seek new information has been elegantly demonstrated in the lab. Volunteers were given a series of difficult decisions to make about the color of shapes on a computer screen. By arranging these shapes in a particular way, the researchers could create conditions in which people *felt* more uncertain about the task but performed no worse. This design nicely isolates the effect a feeling of uncertainty has on our decisions. When asked whether they wanted to see the information again, participants did so only when they felt more uncertain. Just as in the experiments on babies, the participants were relying on internal feelings of uncertainty or confidence to decide whether to ask for help.[18]

Shades of Gray

Being able to track uncertainty is fundamental to how our brains perceive the world. Due to the complexity of our environment and the fact that our senses provide only low-resolution snapshots of our surroundings, we are forced to make assumptions about what is really out there. A powerful approach to solving these inverse problems combines different sources of data according to their reliability or uncertainty. Many aspects of this solution are in keeping with the mathematics of Bayesian inference, although there is a vigorous debate among neuroscientists as to how and whether the brain implements (approximations to) Bayes's rule.[19]

Regardless of how it is done, we can be reasonably sure that computing uncertainty is a fundamental principle of how brains work. If we were unable to represent uncertainty, we would only ever be able to see the world in one particular way (if at all). By representing uncertainty we also acquire our first building block of metacognition—the ability to doubt what our senses are telling us. By itself, the ability to compute uncertainty is not sufficient for full-blown self-awareness. But it is likely sufficient for the rudimentary forms of metacognition that have been discovered in animals and babies. Nabokov's bright line between humans and other species is becoming blurred, with other animals also demonstrating the first signs of metacognitive competence.

But tracking uncertainty is only the beginning of our story. Up until now we have treated the brain as a static perceiver of the world, fixed in place and unable to move around. As soon as we add in the ability to act, we open up entirely new challenges for metacognitive algorithms. Meeting these challenges will require incorporating our next building block: the ability to monitor our actions.

alive becomes very difficult indeed. Consider a humble single-celled bacterium. Living cells depend on managing the acidity of their internal world, because most proteins will cease to function beyond a narrow range of pH. Even simple bacteria have intricate networks of sensors and signaling molecules on their cell surface, which lead to the activation of pumps to restore a neutral pH balance when required.

This is known as homeostasis, and it is ubiquitous in biology. Homeostasis works like the thermostat in your house: when the temperature drops below a certain point, the thermostat switches on the heating, ensuring that the ambience of your living room is kept within a comfortable range. A curious aspect of homeostasis is that it is recursive—it seeks to alter the very same thing that it is monitoring. The thermostat in my living room is trying to regulate the temperature of the same living room, not some room in my neighbor's house. This feature of homeostasis is known as a closed-loop system. If the state it is detecting is in an acceptable range, then all is well. If it's not—if an imbalance in pH or temperature is detected—some action is taken, and the imbalance is corrected. Homeostasis can often be left to its own devices when up and running; it is rare that a corrective action will not have a desired effect, and the control process, while intricate, is computationally simple.

Homeostatic mechanisms, however, operate in the here and now, without caring very much about the future. A simple on-off thermostat cannot "know" that it tends to get colder at night and warmer during the day. It just switches on the heating when the temperature drops below a threshold. In the BBC comedy series *Peep Show*, Jez misunderstands this critical feature of thermostats, telling his housemate Mark, "Let's whack [the boiler] up to 29.... I don't actually want it to be 29, but you've got to give it something to aim for. It'll get hotter, quicker." Mark replies disdainfully (and accurately): "No it won't, it's either on or off. You set it, it achieves the correct temperature, it switches off." You cannot trick a boiler.

The new breed of learning thermostats, such as the Nest, improves on traditional on-off devices by learning the typical rise and fall in temperature over the course of the day and the preferences of the owner for particular temperatures. A smart thermostat can then anticipate when it needs to switch on to

maintain a more even temperature. The reason this is more successful than a good-old-fashioned thermostat is a consequence of a classic proposal in computer science known as the good regulator theorem, which states that the most effective way of controlling a system is to develop an accurate model of that same system. In other words, the more accurate my model of the kind of things that affect the temperature, the more likely I will be able to anticipate when I need to make changes to the heating to keep it within a comfortable range.[1]

The same is true when we move beyond homeostasis to actions that affect the external world. In fact, we can think of all our behavior as a form of elaborate homeostasis, in the sense that many of the things we do are aimed at keeping our internal states within desirable bounds. If I am hungry, I might decide to go and make a sandwich, which makes me feel full again. If I need money to buy ingredients to make a sandwich, I might decide to apply for a job to make money, and so on. This idea—that everything we do in life fits into some grand scheme that serves to minimize the "error" in our internal states—has both its proponents and critics in the field of computational neuroscience. But at least for many of our simpler actions, it provides an elegant framework for thinking about how behavior is monitored and controlled. Let's take a closer look at how this works in practice.[2]

Who Is in Control?

In the same way that there are dedicated sensory parts of the brain —those that handle incoming information from the eyes and ears, for instance—there are also dedicated motor structures that send neural projections down to the spinal cord in order to control and coordinate our muscles. And just as the visual cortex is organized hierarchically, going from input to high-level representations of what is out there in the world, the motor cortex is organized as a descending hierarchy. Regions such as the premotor cortex are involved in creating general plans and intentions (such as "reach to the left"), while lower-level brain areas, such as the primary motor cortex, are left to implement the details. Regions in the prefrontal cortex (PFC) have been suggested to be at the top of

both the perceptual and motor hierarchies. This makes sense if we think of the PFC as being involved in translating high-level perceptual representations (the red ball is over there) into high-level action representations (let's pick up the red ball).[3]

One consequence of the hierarchical organization of action is that when we reach for a cup of coffee, we do not need to consciously activate the sequence of muscles to send our arm and hand out toward the cup. Instead, most action plans are made at a higher level—we want to taste the coffee, and our arm, hand, and mouth coordinate to make it so. This means that in a skilled task such as playing the piano, there is a delicate ballet between conscious plans unfolding further up the hierarchy (choosing how fast to play, or how much emphasis to put on particular passages) and the automatic and unconscious aspects of motor control that send our fingers toward the right keys at just the right time. When watching a concert pianist at work, it seems as though their hands and fingers have a life of their own, while the pianist glides above it all, issuing commands from on high. As the celebrated pianist Vladimir Horowitz declared, "I am a general, my soldiers are the keys." In the more prosaic language of neuroscience, we offload well-learned tasks to unconscious, subordinate levels of action control, intervening only where necessary.[4]

Not all of us can engage in the finger acrobatics required for playing Chopin or Liszt. But many of us regularly engage in a similarly remarkable motor skill on another type of keyboard. I am writing this book on a laptop equipped with a standard QWERTY keyboard, named for the first six letters of the top row. The history of why the QWERTY keyboard, designed by politician and amateur inventor Christopher Latham Sholes in the 1860s, came into being is murky (the earliest typewriters instead had all twenty-six letters of the alphabet organized in a row from A to Z, which its inventors assumed would be the most efficient arrangement). One story is that it was to prevent the early typewriters from getting jammed. Another is that it helped telegraph operators, who received Morse code, quickly transcribe closely related letters in messages. And yet another is that Remington, the first major typewriter manufacturer, wanted to stick with QWERTY to ensure brand loyalty from typists who had trained on its proprietary system.

Whichever theory is correct, the English-speaking world's

QWERTY typewriter has led millions of people to acquire a highly proficient but largely unconscious motor skill. If you are a regular computer user, close your eyes and try to imagine where the letters fall on your keyboard (with the exception of the letters Q-W-E-R-T-Y!). It is not easy, and if you are like me, can only really be done by pretending to type out a word. This neat dissociation between motor skill and conscious awareness makes typing a perfect test bed for studying the different kinds of algorithms involved in unconsciously monitoring and controlling our actions. Typing can also be studied with beautiful precision in the lab: the initiation and timing of keystrokes can be logged by a computer and the movements of people's fingers captured by high-resolution cameras.

Using these methods, the psychologists Gordon Logan and Matthew Crump have carried out detailed and creative experiments to probe how people type. In one of their experiments, people were asked to type out the answers to a classic psychological test, the Stroop task. In the Stroop, people are asked to respond to the color of the ink a word is written in—typing "blue" for blue ink and "red" for red ink, for instance. This is straightforward for most words, but when the words themselves are color words (such as the word "green" written in blue ink, "purple" written in red ink, and so on) it becomes much more difficult, and people slow down and make errors when the word and the ink color don't match. But despite being slower to initiate typing the word, they were no slower to type the letters *within* the word once they had gotten started (for instance, b-l-u-e). This led to the hypothesis that there are multiple action control loops at work: a higher-level loop governing the choice of which word to type, and a lower-level loop that takes this information and works out which keys need to be pressed in which order.[5]

Not only are there multiple levels of action control, but the higher levels know little about the workings of the lower levels. We know this because one of the easiest ways to screw up someone's typing is to ask them to type only the letters in a sentence that would normally be typed by the left (or right) hand. Try sitting at a keyboard and typing only the left-hand letters in the sentence "The cat on the mat" (on a QWERTY keyboard you should produce something like "Tecatteat," depending on whether

you normally hit the space bar with your right or left thumb). It is a fiendishly difficult and frustrating task to assign letters to hands. And yet the lower-level loop controlling our keystrokes does this continuously, at up to seventy words per minute! Some part of us does know the correct hand, but it's not able to get the message out.[6]

Staying the Course

These experiments suggest that fine-scale unconscious adjustments are continuously being made to ensure that our actions stay on track. Occasionally, these unconscious monitoring processes become exposed, similar to how visual illusions revealed the workings of perceptual inference in the previous chapter. For instance, when I commute to work on the London Tube, I have to step onto a series of moving escalators, and I rely on my body making rapid postural adjustments to stop me from falling over when I do so. But this response is so well learned that if the escalator is broken and stationary, it's difficult to stop my motor system from automatically correcting for the impact of the usually moving stairs—so much so that I now have a higher-level expectation that I will stumble slightly going onto a stationary escalator.[7]

In a classic experiment designed to quantify this kind of rapid, automatic error correction, Pierre Fourneret and Marc Jeannerod asked volunteers to move a computer cursor to a target on a screen. By ensuring that participants' hands were hidden (so that they could see only the cursor), the researchers were able to introduce small deviations to the cursor position and observe what happened. They found that when the cursor was knocked off course, people immediately corrected it without being aware of having done so. Their paper concluded: "We found that subjects largely ignored the actual movements that their hand had performed." In other words, a low-level system unconsciously monitors how we are performing the task and corrects—as efficiently as possible—any deviations away from the goal.[8]

One part of the brain that is thought to be critical for supporting these adjustments is known as the cerebellum—from

another button if they detected themselves making an error. Rabbitt precisely measured the time it took for these additional button presses to occur, finding that people were able to correct their own errors very quickly. In fact, they realized they had made an error on average forty milliseconds faster than their fastest responses to external stimuli. This elegant and simple analysis proved that the brain was able to monitor and detect its own errors via an efficient, internal computation, one that did not depend on signals arriving from the outside world.

This rapid process of error detection can lead to an equally rapid process of error correction. In a simple decision about whether a stimulus belongs to category A or B, within only tens of milliseconds after the wrong button is pressed, the muscles controlling the correct response begin to contract in order to rectify the error. And if these corrective processes happen fast enough, they may prevent the error from occurring in the first place. For instance, by the time our muscles are contracting and we are pressing the send button on a rash email, we might have accumulated additional evidence to suggest that this is not a good idea and withhold the critical mouse click at the last moment.[11]

A couple of decades after Rabbitt's experiment, the brain processes that support internal error detection were beginning to be discovered. In his PhD thesis published in 1992, the psychologist William Gehring made electroencephalograph (EEG) recordings from participants while they performed difficult tasks. EEG uses a net of small electrodes to measure the changes in the electrical field outside the head caused by the combined activity of thousands of neurons inside the brain. Gehring found that a unique brain wave was triggered less than one hundred milliseconds after an error was committed. This rapid response helps explain why Rabbitt found that people were often able to very quickly recognize that they had made an error, even before they were told. This activity was labeled the error-related negativity (ERN), which psychologists now affectionately refer to as the "Oh shit!" response.[12]

We now know that the ERN occurs following errors on a multitude of tasks, from pressing buttons to reading aloud, and is generated by a brain region buried in the middle of the frontal lobe: the dorsal anterior cingulate cortex (dACC). This tell-tale

neural signature of self-monitoring is already in place early in human development. In one experiment, twelve-month-old babies were flashed a series of images on a computer screen, and their eye movements recorded. Occasionally one of the images would be a face, and if the babies looked toward it, they would get a reward in the form of music and flashing colored lights. The occasions on which the baby failed to look at the face are errors in the context of the experiment—they did not perform the action that would get them the reward. On these occasions, EEG recordings showed a clear ERN, although somewhat delayed in time compared to what is typically seen in adults.[13]

We can think of the ERN as a special case of a "prediction error" signal. Prediction errors do exactly what they say on the tin—they keep track of errors in our predictions about the future, and they are a central feature of algorithms that can efficiently learn about the world. To see how prediction errors help us learn, imagine that a new coffee shop opens up near your office. You don't yet know how good it is, but they have taken care to buy a top-of-the-line espresso machine and get the ambience just right. Your expectations are high—you predict that the coffee will be good before you've even tasted it. When you sip your first cup, you find that it's not only good—it's one of the best cups of coffee you have had in a long time. The fact that the coffee was better than expected leads you to update your estimate, and it becomes your new favorite stop on the way in to work.

Now let's imagine a few weeks have gone by. The baristas have become complacent and the coffee is no longer as good as it used to be. It might still be good, but compared to what you expected, this is experienced as a negative error in your prediction, and you might feel a little more disappointed than usual.

The ability to make and update predictions depends on a famous brain chemical, dopamine. Dopamine is not only famous, but it is also commonly misunderstood and often referred to as the "pleasure" chemical in the popular media. It is true that dopamine is boosted by things that we enjoy, from money to food to sex. But the idea that dopamine simply signals the rewarding character of an experience is incorrect. In the 1990s, a now classic experiment was carried out by the neuroscientist Wolfram Schultz. He recorded signals from cells in the monkey midbrain that produce

dopamine and deliver it to other brain areas. Schultz trained the monkeys to expect a drop of juice after a light was switched on in the room. Initially, the dopamine cells responded to the juice, consistent with the pleasure theory. But over time, the animals began to learn that the juice was always preceded by the light—they learned to expect the juice—and the dopamine response disappeared.[14]

An elegant explanation for the pattern of dopamine responses in these experiments is that they were tracking the error in the monkeys' prediction about the juice. Early on, the juice was unexpected—just like the unexpectedly good coffee from the new shop. But over time, the monkeys came to expect the juice every time they saw the light, just as we would come to expect good coffee every time we walked into the cafe. Around the same time that Schultz was performing his experiments, the computational neuroscientists Peter Dayan and Read Montague were building on classic work on trial-and-error learning in psychology. A prominent theory, the Rescorla-Wagner rule, proposed that learning should only occur when events are unexpected. This makes intuitive sense: If the coffee is just the same as yesterday, I don't need to alter my estimate of the goodness of the coffee shop. There is no learning to do. Dayan and Montague showed that versions of this algorithm provided an excellent match to the response of dopamine neurons. Shortly after Schultz, Dayan, and Montague's work was published, a series of studies by my former PhD adviser Ray Dolan discovered that the neural response in regions of the human brain that receive dopamine input closely tracks what one would expect of a prediction error signal. Together, these pioneering studies revealed that computing prediction errors and using them to update how we experience the world is a fundamental principle of how brains work.[15]

Now that we're armed with an understanding of prediction errors, we can begin to see how similar computations are important for self-monitoring. Occasionally we directly experience positive or negative feedback about our performance—on an assignment at school, for instance, or when we learn we have beaten our personal best over a half-marathon distance. But in many other areas of everyday life, the feedback may be more subtle, or even absent. One useful way of thinking about the ERN,