Chapter 8

Consciousness

I cannot imagine a consistent theory of everything that ignores consciousness.

Andrei Linde, 2002

We should strive to grow consciousness itself—to generate bigger, brighter lights in an otherwise dark universe.

Giulio Tononi, 2012

We've seen that AI can help us create a wonderful future if we manage to find answers to some of the oldest and toughest problems in philosophy—by the time we need them. We face, in Nick Bostrom's words, philosophy with a deadline. In this chapter, let's explore one of the thorniest philosophical topics of all: consciousness.

Who Cares?

Consciousness is controversial. If you mention the "C-word" to an AI researcher, neuroscientist or psychologist, they may roll their eyes. If they're your mentor, they might instead take pity on you and try to talk you out of wasting your time on what they consider a hopeless and unscientific problem. Indeed, my friend Christof Koch, a renowned neuroscientist who leads the Allen Institute for Brain Science, told me that he was once warned of working on consciousness before he had tenure—by none less than Nobel laureate Francis Crick. If you look up "consciousness" in the 1989 *Macmillan Dictionary of Psychology*, you're informed that "Nothing worth reading has been written on it."[1] As I'll explain in this chapter, I'm more optimistic!

Although thinkers have pondered the mystery of consciousness for thousands of years, the rise of AI adds a sudden urgency, in particular to the question of predicting which intelligent entities have subjective experiences. As we saw in chapter 3, the question of whether intelligent machines should be granted some form of rights depends crucially on whether they're conscious and can suffer or feel joy. As we discussed in chapter 7, it becomes hopeless to formulate utilitarian ethics based on maximizing positive experiences without knowing which intelligent entities are capable of having them. As mentioned in chapter 5, some people might prefer their robots to be unconscious to avoid feeling slave-owner guilt. On the other hand, they may desire the opposite if they upload their minds to break free from biological limitations: after all, what's the point of uploading yourself into a robot that talks and acts like you if it's a mere unconscious zombie, by which I mean that being the uploaded you doesn't feel like anything? Isn't this equivalent to committing suicide from your subjective point of view, even though your friends may not realize that your subjective experience has died?

For the long-term cosmic future of life (chapter 6), understanding what's conscious and what's not becomes pivotal: if technology enables intelligent life to flourish throughout our Universe for billions of years, how can we be sure that this life is conscious and able to appreciate what's happening? If not, then would it be, in the words of the famous physicist Erwin Schrödinger, "a play before empty benches, not existing for anybody, thus quite properly speaking not

existing”?[2] In other words, if we enable high-tech descendants that we mistakenly think are conscious, would this be the ultimate zombie apocalypse, transforming our grand cosmic endowment into nothing but an astronomical waste of space?

What Is Consciousness?

Many arguments about consciousness generate more heat than light because the antagonists are talking past each other, unaware that they're using different definitions of the C-word. Just as with "life" and "intelligence," there's no undisputed correct definition of the word "consciousness." Instead, there are many competing ones, including sentience, wakefulness, self-awareness, access to sensory input and ability to fuse information into a narrative.[3] In our exploration of the future of intelligence, we want to take a maximally broad and inclusive view, not limited to the sorts of biological consciousness that exist so far. That's why the definition I gave in chapter 1, which I'm sticking with throughout this book, is very broad:

> consciousness = *subjective experience*

In other words, if it feels like something to be you right now, then you're conscious. It's this particular definition of consciousness that gets to the crux of all the AI-motivated questions in the previous section: Does it feel like something to be Prometheus, AlphaGo or a self-driving Tesla?

To appreciate how broad our consciousness definition is, note that it doesn't mention behavior, perception, self-awareness, emotions or attention. So by this definition, you're conscious also when you're dreaming, even though you lack wakefulness or access to sensory input and (hopefully!) aren't sleepwalking and doing things. Similarly, any system that experiences pain is conscious in this sense, even if it can't move. Our definition leaves open the possibility that some future AI systems may be conscious too, even if they exist merely as software and aren't connected to sensors or robotic bodies.

With this definition, it's hard not to care about consciousness. As Yuval Noah Harari puts it in his book *Homo Deus:*[4] "If any scientist wants to argue that subjective experiences are irrelevant, their challenge is to explain why torture or

rape are wrong without reference to any subjective experience." Without such reference, it's all just a bunch of elementary particles moving around according to the laws of physics—and what's wrong with that?

What's the Problem?

So what precisely is it that we don't understand about consciousness? Few have thought harder about this question than David Chalmers, a famous Australian philosopher rarely seen without a playful smile and a black leather jacket—which my wife liked so much that she gave me a similar one for Christmas. He followed his heart into philosophy despite making the finals at the International Mathematics Olympiad—and despite the fact that his only B grade in college, shattering his otherwise straight As, was for an introductory philosophy course. Indeed, he seems utterly undeterred by put-downs or controversy, and I've been astonished by his ability to politely listen to uninformed and misguided criticism of his own work without even feeling the need to respond.

As David has emphasized, there are really two separate mysteries of the mind. First, there's the mystery of how a brain processes information, which David calls the "easy" problems. For example, how does a brain attend to, interpret and respond to sensory input? How can it report on its internal state using language? Although these questions are actually extremely difficult, they're by our definitions not mysteries of consciousness, but mysteries of intelligence: they ask how a brain remembers, computes and learns. Moreover, we saw in the first part of the book how AI researchers have started to make serious progress on solving many of these "easy problems" with machines—from playing Go to driving cars, analyzing images and processing natural language.

Then there's the separate mystery of why you have a subjective experience, which David calls the *hard* problem. When you're driving, you're experiencing colors, sounds, emotions, and a feeling of self. But why are you experiencing anything at all? Does a self-driving car experience anything at all? If you're racing against a self-driving car, you're both inputting information from sensors, processing it and outputting motor commands. But subjectively *experiencing* driving is something logically separate—is it optional, and if so, what causes it?

I approach this hard problem of consciousness from a physics point of view. From my perspective, a conscious person is simply food, rearranged. So why is one arrangement conscious, but not the other? Moreover, physics teaches us that food is simply a large number of quarks and electrons, arranged in a certain way. So which particle arrangements are conscious and which aren't?[*1]

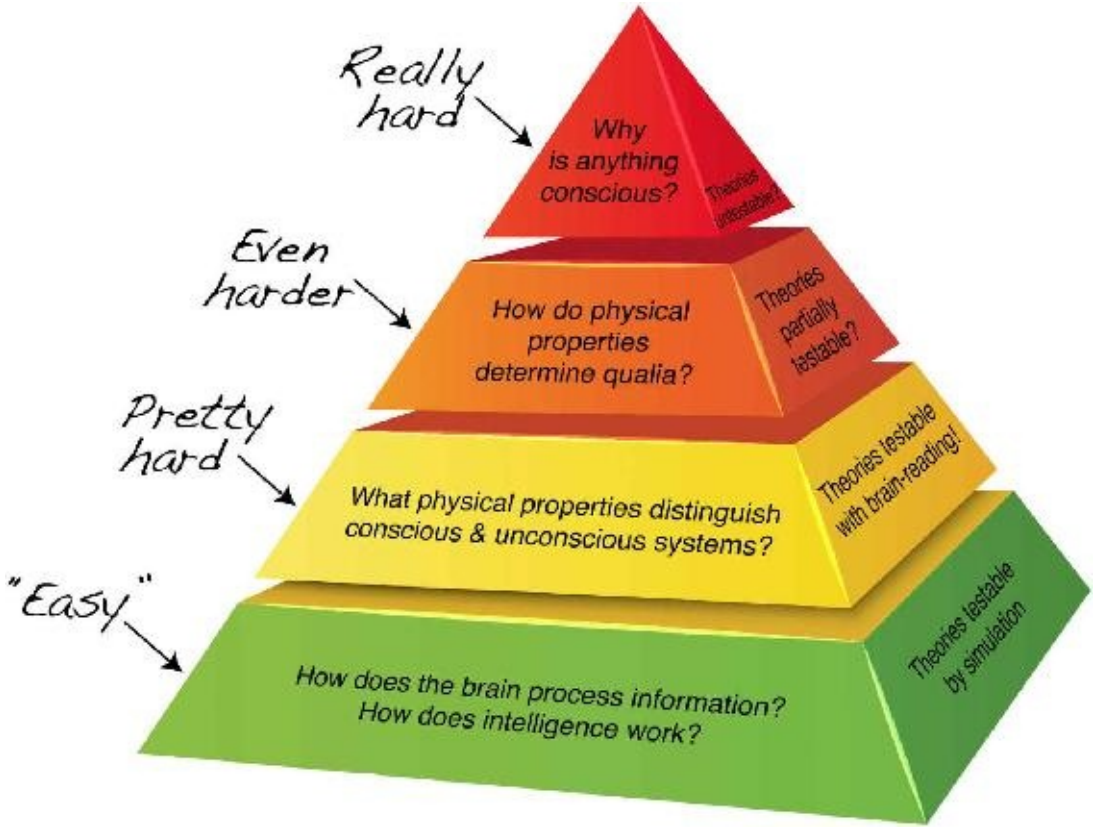

Figure 8.1: Understanding the mind involves a hierarchy of problems. What David Chalmers calls the "easy" problems can be posed without mentioning subjective experience. The apparent fact that some but not all physical systems are conscious poses three separate questions. If we have a theory for answering the question that defines the "pretty hard problem," then it can be experimentally tested. If it works, then we can build on it to tackle the tougher questions above.

What I like about this physics perspective is that it transforms the hard problem that we as humans have struggled with for millennia into a more focused version that's easier to tackle with the methods of science. Instead of starting with a hard *problem* of why an arrangement of particles can feel conscious, let's start with a hard *fact* that some arrangements of particles do feel conscious while others don't. For example, you know that the particles that make up your brain are in a conscious arrangement right now, but not when you're in deep dreamless sleep.

This physics perspective leads to three separate hard questions about consciousness, as shown in figure 8.1. First of all, what properties of the particle

arrangement make the difference? Specifically, what physical properties distinguish conscious and unconscious systems? If we can answer that, then we can figure out which AI systems are conscious. In the more immediate future, it can also help emergency-room doctors determine which unresponsive patients are conscious.

Second, how do physical properties determine what the experience is like? Specifically, what determines *qualia*, basic building blocks of consciousness such as the redness of a rose, the sound of a cymbal, the smell of a steak, the taste of a tangerine or the pain of a pinprick?[*2]

Third, why is anything conscious? In other words, is there some deep undiscovered explanation for why clumps of matter can be conscious, or is this just an unexplainable brute fact about the way the world works?

The computer scientist Scott Aaronson, a former MIT colleague of mine, has lightheartedly called the first question the "pretty hard problem" (PHP), as has David Chalmers. In that spirit, let's call the other two the "even harder problem" (EHP) and the "really hard problem" (RHP), as illustrated in figure 8.1.[*3]

Is Consciousness Beyond Science?

When people tell me that consciousness research is a hopeless waste of time, the main argument they give is that it's "unscientific" and always will be. But is that really true? The influential Austro-British philosopher Karl Popper popularized the now widely accepted adage "If it's not falsifiable, it's not scientific." In other words, science is all about testing theories against observations: if a theory can't be tested even in principle, then it's logically impossible to ever falsify it, which by Popper's definition means that it's unscientific.

So could there be a scientific theory that answers any of the three consciousness questions in figure 8.1? Please let me try to persuade you that the answer is a resounding YES!, at least for the pretty hard problem: "What physical properties distinguish conscious and unconscious systems?" Suppose that someone has a theory that, given any physical system, answers the question of whether the system is conscious with "yes," "no" or "unsure." Let's hook your brain up to a device that measures some of the information processing in different parts of your brain, and let's feed this information into a computer program that uses the consciousness theory to predict which parts of that information are conscious, and presents you with its predictions in real time on a screen, as in figure 8.2. First you think of an apple. The screen informs you that there's information about an apple in your brain which you're aware of, but that there's also information in your brainstem about your pulse that you're unaware of. Would you be impressed? Although the first two predictions of the theory were correct, you decide to do some more rigorous testing. You think about your mother and the computer informs you that there's information in your brain about your mother but that you're unaware of this. The theory made an incorrect prediction, which means that it's ruled out and goes in the garbage dump of scientific history together with Aristotelian mechanics, the luminiferous aether, geocentric cosmology and countless other failed ideas. Here's the key point: Although the theory was wrong, it was *scientific*! Had it not been scientific, you wouldn't have been able to test it and rule it out.

Someone might criticize this conclusion and say that *they* have no evidence of what you're conscious of, or even of you being conscious at all: although they heard you say that you're conscious, an unconscious zombie could conceivably

say the same thing. But this doesn't make that consciousness theory unscientific, because they can trade places with you and test whether it correctly predicts *their own* conscious experiences.

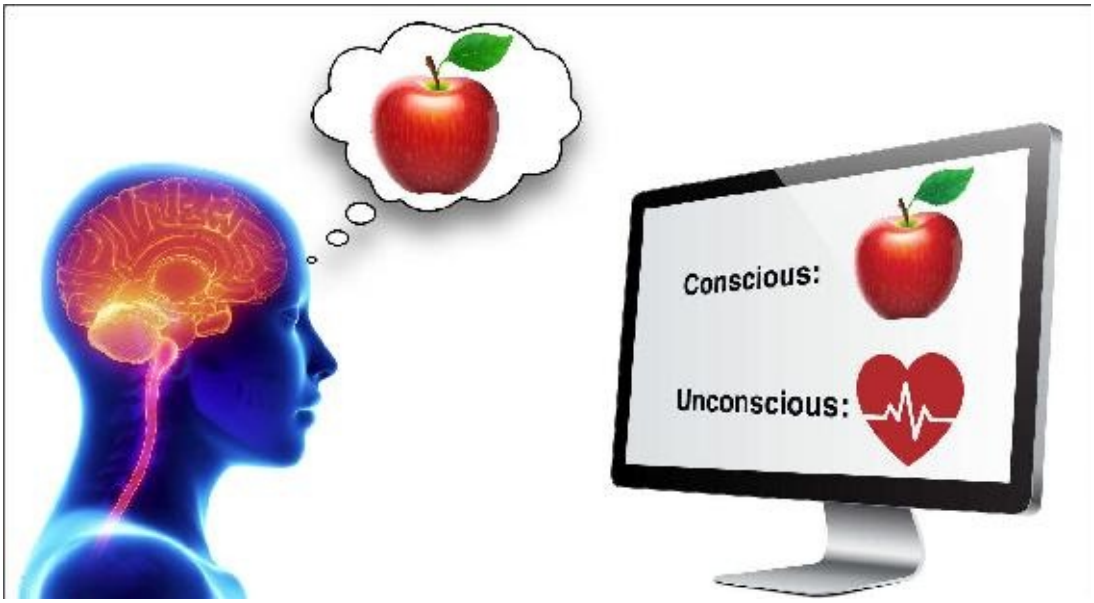

Figure 8.2: Suppose that a computer measures information being processed in your brain and predicts which parts of it you're aware of according to a theory of consciousness. You can scientifically test this theory by checking whether its predictions are correct, matching your subjective experience.

On the other hand, if the theory refuses to make any predictions, merely replying "unsure" whenever queried, then it's untestable and hence unscientific. This might happen because it's applicable only in some situations, because the required computations are too hard to carry out in practice or because the brain sensors are no good. Today's most popular scientific theories tend to be somewhere in the middle, giving testable answers to some but not all of our questions. For example, our core theory of physics will refuse to answer questions about systems that are simultaneously extremely small (requiring quantum mechanics) and extremely heavy (requiring general relativity), because we haven't yet figured out which mathematical equations to use in this case. This core theory will also refuse to predict the exact masses of all possible atoms —in this case, we think we have the necessary equations, but we haven't managed to accurately compute their solutions. The more dangerously a theory lives by sticking its neck out and making testable predictions, the more useful it is, and the more seriously we take it if it survives all our attempts to kill it. Yes, we can only test *some* predictions of consciousness theories, but that's how it is

for *all* physical theories. So let's not waste time whining about what we can't test, but get to work testing what we *can* test!

In summary, any theory predicting which physical systems are conscious (the pretty hard problem) is scientific, as long as it can predict which of your brain processes are conscious. However, the testability issue becomes less clear for the higher-up questions in figure 8.1. What would it mean for a theory to predict how you subjectively experience the color red? And if a theory purports to explain why there is such a thing as consciousness in the first place, then how do you test it experimentally? Just because these questions are hard doesn't mean that we should avoid them, and we'll indeed return to them below. But when confronted with several related unanswered questions, I think it's wise to tackle the easiest one first. For this reason, my consciousness research at MIT is focused squarely on the base of the pyramid in figure 8.1. I recently discussed this strategy with my fellow physicist Piet Hut from Princeton, who joked that trying to build the top of the pyramid before the base would be like worrying about the interpretation of quantum mechanics before discovering the Schrödinger equation, the mathematical foundation that lets us predict the outcomes of our experiments.

When discussing what's beyond science, it's important to remember that the answer depends on time! Four centuries ago, Galileo Galilei was so impressed by math-based physics theories that he described nature as "a book written in the language of mathematics." If he threw a grape and a hazelnut, he could accurately predict the shapes of their trajectories and when they would hit the ground. Yet he had no clue why one was green and the other brown, or why one was soft and the other hard—these aspects of the world were beyond the reach of science at the time. But not forever! When James Clerk Maxwell discovered his eponymous equations in 1861, it became clear that light and colors could also be understood mathematically. We now know that the aforementioned Schrödinger equation, discovered in 1925, can be used to predict all properties of matter, including what's soft or hard. While theoretical progress has enabled ever more scientific predictions, technological progress has enabled ever more experimental tests: almost everything we now study with telescopes, microscopes or particle colliders was once beyond science. In other words, the purview of science has expanded dramatically since Galileo's days, from a tiny fraction of all phenomena to a large percentage, including subatomic particles, black holes and our cosmic origins 13.8 billion years ago. This raises the question: What's left?

To me, consciousness is the elephant in the room. Not only do you know that you're conscious, but it's *all* you know with complete certainty—everything else is inference, as René Descartes pointed out back in Galileo's time. Will theoretical and technological progress eventually bring even consciousness firmly into the domain of science? We don't know, just as Galileo didn't know whether we'd one day understand light and matter.[*4] Only one thing is guaranteed: we won't succeed if we don't try! That's why I and many other scientists around the world are trying hard to formulate and test theories of consciousness.

Experimental Clues About Consciousness

Lots of information processing is taking place in our heads right now. Which of it is conscious and which isn’t? Before exploring consciousness theories and what they predict, let’s look at what experiments have taught us so far, ranging from traditional low-tech or no-tech observations to state-of-the-art brain measurements.

What Behaviors Are Conscious?

If you multiply 32 by 17 in your head, you're conscious of many of the inner workings of your computation. But suppose I instead show you a portrait of Albert Einstein and tell you to say the name of its subject. As we saw in chapter 2, this too is a computational task: your brain is evaluating a function whose input is information from your eyes about a large number of pixel colors and whose output is information to muscles controlling your mouth and vocal cords. Computer scientists call this task "image classification" followed by "speech synthesis." Although this computation is way more complicated than your multiplication task, you can do it much faster, seemingly without effort, and without being conscious of the details of *how* you do it. Your subjective experience consists merely of looking at the picture, experiencing a feeling of recognition and hearing yourself say "Einstein."

Psychologists have long known that you can unconsciously perform a wide range of other tasks and behaviors as well, from blink reflexes to breathing, reaching, grabbing and keeping your balance. Typically, you're consciously aware of what you did, but not how you did it. On the other hand, behaviors that involve unfamiliar situations, self-control, complicated logical rules, abstract reasoning or manipulation of language tend to be conscious. They're known as *behavioral correlates of consciousness*, and they're closely linked to the effortful, slow and controlled way of thinking that psychologists call "System 2."[5]

It's also known that you can convert many routines from conscious to unconscious through extensive practice, for example walking, swimming, bicycling, driving, typing, shaving, shoe tying, computer-gaming and piano playing.[6] Indeed, it's well known that experts do their specialties best when they're in a state of "flow," aware only of what's happening at a higher level, and unconscious of the low-level details of how they're doing it. For example, try reading the next sentence while being consciously aware of every single letter, as when you first learned to read. Can you feel how much slower it is, compared to when you're merely conscious of the text at the level of words or ideas?

Indeed, unconscious information processing appears not only to be possible, but also to be more the rule than the exception. Evidence suggests that of the

roughly 10^7 bits of information that enter our brain each second from our sensory organs, we can be aware only of a tiny fraction, with estimates ranging from 10 to 50 bits.[7] This suggests that the information processing that we're consciously aware of is merely the tip of the iceberg.

Taken together, these clues have led some researchers to suggest that conscious information processing should be thought of as the CEO of our mind, dealing with only the most important decisions requiring complex analysis of data from all over the brain.[8] This would explain why, just like the CEO of a company, it usually doesn't want to be distracted by knowing everything its underlings are up to—but it can find them out if desired. To experience this selective attention in action, look at that word "desired" again: fix your gaze on the dot over the "i" and, without moving your eyes, shift your attention from the dot to the whole letter and then to the whole word. Although the information from your retina stayed the same, your conscious experience changed. The CEO metaphor also explains why expertise becomes unconscious: after painstakingly figuring out how to read and type, the CEO delegates these routine tasks to unconscious subordinates to be able to focus on new higher-level challenges.

Where Is Consciousness?

Clever experiments and analyses have suggested that consciousness is limited not merely to certain behaviors, but also to certain parts of the brain. Which are the prime suspects? Many of the first clues came from patients with brain lesions: localized brain damage caused by accidents, strokes, tumors or infections. But this was often inconclusive. For example, does the fact that lesions in the back of the brain can cause blindness mean that this is the site of visual consciousness, or does it merely mean that visual information passes through there en route to wherever it will later become conscious, just as it first passes through the eyes?

Although lesions and medical interventions haven't pinpointed the locations of conscious experiences, they've helped narrow down the options. For example, I know that although I experience pain in my hand as actually occurring there, the pain experience must occur elsewhere, because a surgeon once switched off my hand pain without doing anything to my hand: he merely anesthetized nerves in my shoulder. Moreover, some amputees experience phantom pain that feels as though it's in their nonexistent hand. As another example, I once noticed that when I looked only with my right eye, part of my visual field was missing—a doctor determined that my retina was coming loose and reattached it. In contrast, patients with certain brain lesions experience *hemineglect,* where they too miss information from half their visual field, but aren't even aware that it's missing—for example, failing to notice and eat the food on the left half of their plate. It's as if consciousness about half of their world has disappeared. But are those damaged brain areas supposed to generate the spatial experience, or were they merely feeding spatial information to the sites of consciousness, just as my retina did?

The pioneering U.S.-Canadian neurosurgeon Wilder Penfield found in the 1930s that his neurosurgery patients reported different parts of their body being touched when he electrically stimulated specific brain areas in what's now called the *somatosensory cortex* (figure 8.3).[9] He also found that they involuntarily moved different parts of their body when he stimulated brain areas in what's now called the *motor cortex.* But does that mean that information processing in these brain areas corresponds to consciousness of touch and motion?

Fortunately, modern technology is now giving us much more detailed clues.

Although we're still nowhere near being able to measure every single firing of all of your roughly hundred billion neurons, brain-reading technology is advancing rapidly, involving techniques with intimidating names such as fMRI, EEG, MEG, ECoG, ePhys and fluorescent voltage sensing. fMRI, which stands for functional magnetic resonance imaging, measures the magnetic properties of hydrogen nuclei to make a 3-D map of your brain roughly every second, with millimeter resolution. EEG (electroencephalography) and MEG (magnetoencephalography) measure the electric and magnetic field outside your head to map your brain thousands of times per second, but with poor resolution, unable to distinguish features smaller than a few centimeters. If you're squeamish, you'll appreciate that these three techniques are all noninvasive. If you don't mind opening up your skull, you have additional options. ECoG (electrocorticography) involves placing say a hundred wires on the surface of your brain, while ePhys (electrophysiology) involves inserting microwires, which are sometimes thinner than a human hair, deep into the brain to record voltages from as many as a thousand simultaneous locations. Many epileptic patients spend days in the hospital while ECoG is used to figure out what part of their brain is triggering seizures and should be resected, and kindly agree to let neuroscientists perform consciousness experiments on them in the meantime. Finally, fluorescent voltage sensing involves genetically manipulating neurons to emit flashes of light when firing, enabling their activity to be measured with a microscope. Out of all the techniques, it has the potential to rapidly monitor the largest number of neurons, at least in animals with transparent brains—such as the *C. elegans* worm with its 302 neurons and the larval zebrafish with its about 100,000.

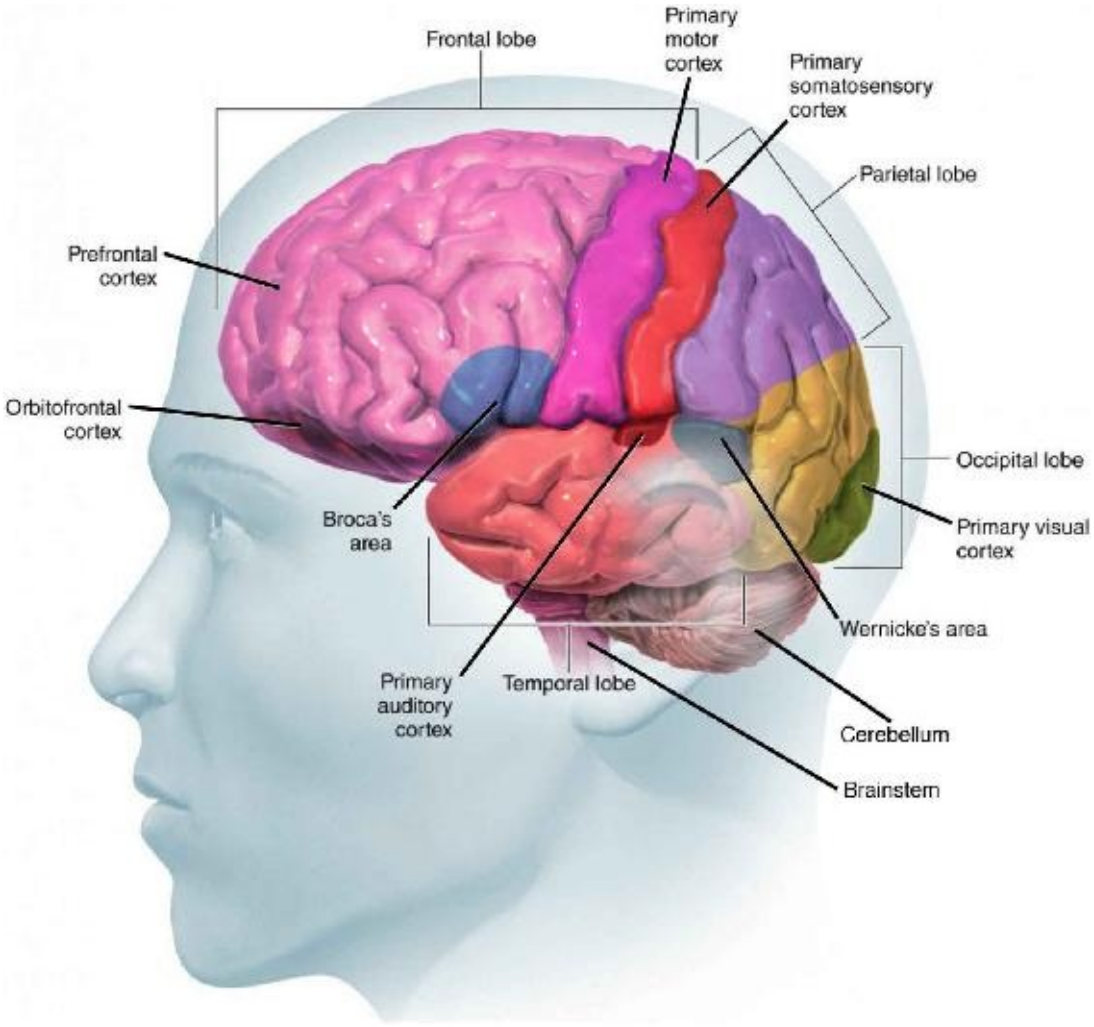

Figure 8.3: The visual, auditory, somatosensory and motor cortices are involved with vision, hearing, the sense of touch and motion activation, respectively—but that doesn't prove they're where *consciousness* of vision, hearing, touch and motion occurs. Indeed, recent research suggests that the primary visual cortex is completely unconscious, together with the cerebellum and brainstem. Image courtesy of Lachina (www.lachina.com).

Although Francis Crick warned Christof Koch about studying consciousness, Christof refused to give up and and eventually won Francis over. In 1990, they wrote a seminal paper about what they called "neural correlates of consciousness" (NCCs), asking which specific brain processes corresponded to conscious experiences. For thousands of years, thinkers had had access to the information processing in their brains only via their subjective experience and

behavior. Crick and Koch pointed out that brain-reading technology was suddenly providing independent access to this information, allowing scientific study of which information processing corresponded to what conscious experience. Sure enough, technology-driven measurements have by now turned the quest for NCCs into quite a mainstream part of neuroscience, one whose thousands of publications extend into even the most prestigious journals.[10]

What are the conclusions so far? To get a flavor for NCC detective work, let's first ask whether your retina is conscious, or whether it's merely a zombie system that records visual information, processes it and sends it on to a system downstream in your brain where your subjective visual experience occurs. In the left panel of figure 8.4, which square is darker: the one labeled A or B? A, right? No, they're in fact identically colored, which you can verify by looking at them through small holes between your fingers. This proves that your visual experience can't reside entirely in your retina, since if it did, they'd look the same.

Now look at the right panel of figure 8.4. Do you see two women or a vase? If you look long enough, you'll subjectively experience both in succession, even though the information reaching your retina remains the same. By measuring what happens in your brain during the two situations, one can tease apart what makes the difference—and it's not the retina, which behaves identically in both cases.

The death blow to the conscious-retina hypothesis comes from a technique called "continuous flash suppression" pioneered by Christof Koch, Stanislas Dehaene and collaborators: it's been discovered that if you make one of your eyes watch a complicated sequence of rapidly changing patterns, then this will distract your visual system to such an extent that you'll be completely unaware of a still image shown to the other eye.[11] In summary, you can have a visual image in your retina without experiencing it, and you can (while dreaming) experience an image without it being on your retina. This proves that your two retinas don't host your visual consciousness any more than a video camera does, even though they perform complicated computations involving over a hundred million neurons.

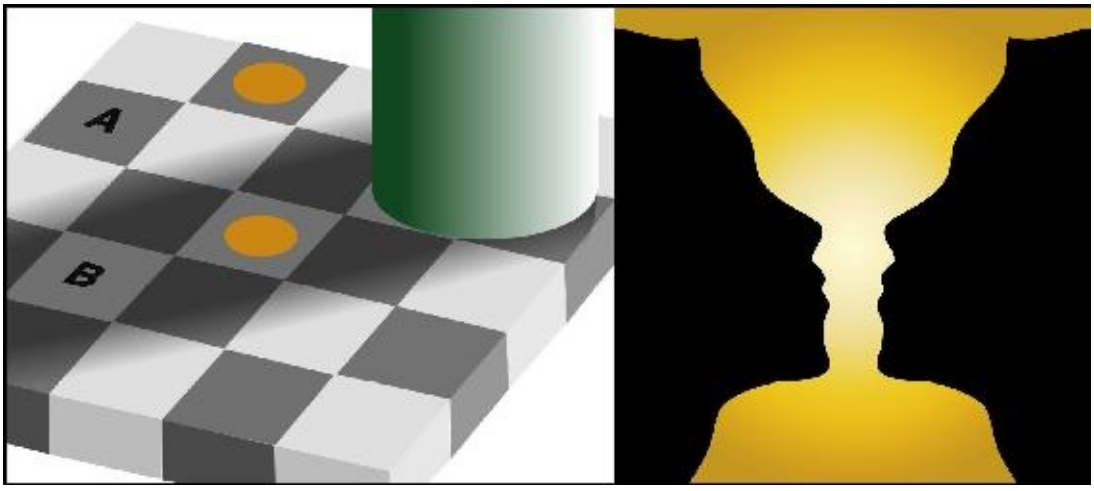

Figure 8.4: Which square is darker—A or B? What do you see on the right—a vase, two women or both in succession? Illusions such as these prove that your visual consciousness can't be in your eyes or other early stages of your visual system, because it doesn't depend only on what's in the picture.

NCC researchers also use continuous flash suppression, unstable visual/auditory illusions and other tricks to pinpoint which of your brain regions *are* responsible for each of your conscious experiences. The basic strategy is to compare what your neurons are doing in two situations where essentially everything (including your sensory input) is the same—except your conscious experience. The parts of your brain that are measured to behave differently are then identified as NCCs.

Such NCC research has proven that none of your consciousness resides in your gut, even though that's the location of your enteric nervous system with its whopping half-billion neurons that compute how to optimally digest your food; feelings such as hunger and nausea are instead produced in your brain. Similarly, none of your consciousness appears to reside in the brainstem, the bottom part of the brain that connects to the spinal cord and controls breathing, heart rate and blood pressure. More shockingly, your consciousness doesn't appear to extend to your cerebellum (figure 8.3), which contains about two-thirds of all your neurons: patients whose cerebellum is destroyed experience slurred speech and clumsy motion reminiscent of a drunkard, but remain fully conscious.

The question of which parts of your brain *are* responsible for consciousness

remains open and controversial. Some recent NCC research suggests that your consciousness mainly resides in a “hot zone” involving the thalamus (near the middle of your brain) and the rear part of the cortex (the outer brain layer consisting of a crumpled-up six-layer sheet which, if flattened out, would have the area of a large dinner napkin).[12] This same research controversially suggests that the primary visual cortex at the very back of the head is an exception to this, being as unconscious as your eyeballs and your retinas.

When Is Consciousness?

So far, we've looked at experimental clues regarding what types of information processing are conscious and where consciousness occurs. But *when* does it occur? When I was a kid, I used to think that we become conscious of events as they happen, with absolutely no time lag or delay. Although that's still how it subjectively feels to me, it clearly can't be correct, since it takes time for my brain to process the information that enters via my sensory organs. NCC researchers have carefully measured how long, and Christof Koch's summary is that it takes about a quarter of a second from when light enters your eye from a complex object until you consciously perceive seeing it as what it is.[13] This means that if you're driving down a highway at fifty-five miles per hour and suddenly see a squirrel a few meters in front of you, it's too late for you to do anything about it, because you've already run over it!

In summary, your consiousness lives in the past, with Christof Koch estimating that it lags behind the outside world by about a quarter second. Intriguingly, you can often react to things faster than you can become conscious of them, which proves that the information processing in charge of your most rapid reactions must be unconscious. For example, if a foreign object approaches your eye, your blink reflex can close your eyelid within a mere tenth of a second. It's as if one of your brain systems receives ominous information from the visual system, computes that your eye is in danger of getting struck, emails your eye muscles instructions to blink and simultaneously emails the conscious part of your brain saying "Hey, we're going to blink." By the time this email has been read and included into your conscious experience, the blink has already happened.

Indeed, the system that reads that email is continually bombarded with messages from all over your body, some more delayed than others. It takes longer for nerve signals to reach your brain from your fingers than from your face because of distance, and it takes longer for you to analyze images than sounds because it's more complicated—which is why Olympic races are started with a bang rather than with a visual cue. Yet if you touch your nose, you consciously experience the sensation on your nose and fingertip as simultaneous, and if you clap your hands, you see, hear and feel the clap at exactly the same time.[14] This means that your full conscious experience of an event isn't created

until the last slowpoke email reports have trickled in and been analyzed.

A famous family of NCC experiments pioneered by physiologist Benjamin Libet has shown that the sort of actions you can perform unconsciously aren't limited to rapid responses such as blinks and ping-pong smashes, but also include certain decisions that you might attribute to free will—brain measurements can sometimes predict your decision before you become conscious of having made it.[15]

Theories of Consciousness

We've just seen that, although we still don't understand consciousness, we have amazing amounts of experimental data about various aspects of it. But all this data comes from *brains*, so how can it teach us anything about consciousness in *machines*? This requires a major extrapolation beyond our current experimental domain. In other words, it requires a *theory.*

Why a Theory?

To appreciate why, let's compare theories of consciousness with theories of gravity. Scientists started taking Newton's theory of gravity seriously because they got more out of it than they put into it: simple equations that fit on a napkin could accurately predict the outcome of every gravity experiment ever conducted. They therefore also took seriously its predictions far beyond the domain where it had been tested, and these bold extrapolations turned out to work even for the motions of galaxies in clusters millions of light-years across. However, the predictions were off by a tiny amount for the motion of Mercury around the Sun. Scientists then started taking seriously Einstein's improved theory of gravity, general relativity, because it was arguably even more elegant and economical, and correctly predicted even what Newton's theory got wrong. They consequently took seriously also its predictions far beyond the domain where it had been tested, for phenomena as exotic as black holes, gravitational waves in the very fabric of spacetime, and the expansion of our Universe from a hot fiery origin—all of which were subsequently confirmed by experiment.

Analogously, if a mathematical theory of consciousness whose equations fit on a napkin could successfully predict the outcomes of all experiments we perform on brains, then we'd start taking seriously not merely the theory itself, but also its predictions for consciousness beyond brains—for example, in machines.

Consciousness from a Physics Perspective

Although some theories of consciousness date back to antiquity, most modern ones are grounded in neuropsychology and neuroscience, attempting to explain and predict consciousness in terms of neural events occurring in the brain.[16] Although these theories have made some successful predictions for neural correlates of consciousness, they neither can nor aspire to make predictions about machine consciousness. To make the leap from brains to machines, we need to generalize from NCCs to PCCs: *physical correlates of consciousness*, defined as the patterns of moving particles that are conscious. Because if a theory can correctly predict what's conscious and what's not by referring only to physical building blocks such as elementary particles and force fields, then it can make predictions not merely for brains, but also for any other arrangements of matter, including future AI systems. So let's take a physics perspective: What particle arrangements are conscious?

But this really raises another question: How can something as complex as consciousness be made of something as simple as particles? I think it's because it's a phenomenon that has properties above and beyond those of its particles. In physics, we call such phenomena "emergent."[17] Let's understand this by looking at an emergent phenomenon that's simpler than consciousness: wetness.

A drop of water is wet, but an ice crystal and a cloud of steam aren't, even though they're made of identical water molecules. Why? Because the property of wetness depends only on the arrangement of the molecules. It makes absolutely no sense to say that a single water molecule is wet, because the phenomenon of wetness emerges only when there are many molecules, arranged in the pattern we call liquid. So solids, liquids and gases are all emergent phenomena: they're more than the sum of their parts, because they have properties above and beyond the properties of their particles. They have properties that their particles lack.

Now just like solids, liquids and gases, I think consciousness is an emergent phenomenon, with properties above and beyond those of its particles. For example, entering deep sleep extinguishes consciousness, by merely rearranging the particles. In the same way, my consciousness would disappear if I froze to death, which would rearrange my particles in a more unfortunate way.

When you put lots of particles together to make anything from water to a

brain, new phenomena with observable properties emerge. We physicists love studying these emergent properties, which can often be identified by a small set of numbers that you can go out and measure—quantities such as how viscous the substance is, how compressible it is and so on. For example, if a substance is so viscous that it's rigid, we call it a solid, otherwise we call it a fluid. And if a fluid isn't compressible, we call it a liquid, otherwise we call it a gas or a plasma, depending on how well it conducts electricity.

Consciousness as Information

So could there be analogous quantities that quantify consciousness? The Italian neuroscientist Giulio Tononi has proposed one such quantity, which he calls the *"integrated information,"* denoted by the Greek letter Φ (*Phi*), which basically measures how much different parts of a system know about each other (see figure 8.5).

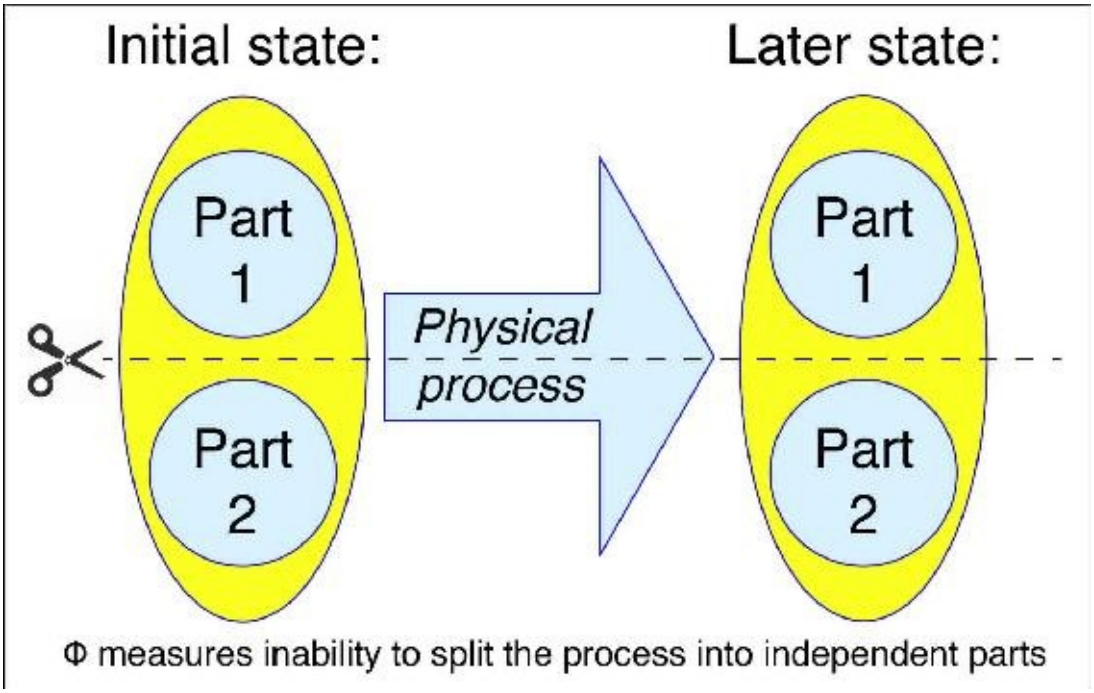

Figure 8.5: Given a physical process that, with the passage of time, transforms the initial state of a system into a new state, its *integrated information* Φ measures inability to split the process into independent parts. If the future state of each part depends only on its own past, not on what the other part has been doing, then Φ = 0: what we called one system is really two independent systems that don't communicate with each other at all.

I first met Giulio at a 2014 physics conference in Puerto Rico to which I'd invited him and Christof Koch, and he struck me as the ultimate renaissance man who'd have blended right in with Galileo and Leonardo da Vinci. His quiet demeanor couldn't hide his incredible knowledge of art, literature and philosophy, and his culinary reputation preceded him: a cosmopolitan TV journalist had recently told me how Giulio had, in just a few minutes, whipped up the most delicious salad he'd tasted in his life. I soon realized that behind his soft-spoken demeanor was a fearless intellect who'd follow the evidence wherever it took him, regardless of the preconceptions and taboos of the establishment. Just as Galileo had pursued his mathematical theory of motion despite establishment pressure not to challenge geocentrism, Giulio had developed the most mathematically precise consciousness theory to date, *integrated information theory* (IIT).

I'd been arguing for decades that consciousness is the way information feels when being processed in certain complex ways.[18] IIT agrees with this and replaces my vague phrase "certain complex ways" by a precise definition: the information processing needs to be integrated, that is, Φ needs to be large. Giulio's argument for this is as powerful as it is simple: the conscious system needs to be integrated into a unified whole, because if it instead consisted of two independent parts, then they'd feel like two separate conscious entities rather than one. In other words, if a conscious part of a brain or computer can't communicate with the rest, then the rest can't be part of its subjective experience.

Giulio and his collaborators have measured a simplified version of Φ by using EEG to measure the brain's response to magnetic stimulation. Their "consciousness detector" works really well: it determined that patients were conscious when they were awake or dreaming, but unconscious when they were anesthetized or in deep sleep. It even discovered consciousness in two patients suffering from "locked-in" syndrome, who couldn't move or communicate in any normal way.[19] So this is emerging as a promising technology for doctors in the future to figure out whether certain patients are conscious or not.

Anchoring Consciousness in Physics

IIT is defined only for discrete systems that can be in a finite number of states, for example bits in a computer memory or oversimplified neurons that can be either on or off. This unfortunately means that IIT isn't defined for most traditional physical systems, which can change continuously—for example, the position of a particle or the strength of a magnetic field can take any of an infinite number of values.[20] If you try to apply the IIT formula to such systems, you'll typically get the unhelpful result that Φ is infinite. Quantum-mechanical systems can be discrete, but the original IIT isn't defined for quantum systems. So how can we anchor IIT and other information-based consciousness theories on a solid physical foundation?

We can do this by building on what we learned in chapter 2 about how clumps of matter can have emergent properties that are related to information. We saw that for something to be usable as a memory device that can store information, it needs to have many long-lived states. We also saw that being *computronium*, a substance that can do computations, in addition requires complex dynamics: the laws of physics need to make it change in ways that are complicated enough to be able to implement arbitrary information processing. Finally, we saw how a neural network, for example, is a powerful substrate for learning because, simply by obeying the laws of physics, it can rearrange itself to get better and better at implementing desired computations. Now we're asking an additional question: What makes a blob of matter able to have a subjective experience? In other words, under what conditions will a blob of matter be able to do these four things?

1. remember
2. compute
3. learn
4. experience

We explored the first three in chapter 2, and are now tackling the fourth. Just as Margolus and Toffoli coined the term *computronium* for a substance that can perform arbitrary computations, I like to use the term *sentronium* for the most

general substance that has subjective experience (is sentient).[*5]

But how can consciousness feel so non-physical if it's in fact a physical phenomenon? How can it feel so independent of its physical substrate? I think it's because it *is* rather independent of its physical substrate, the stuff in which it is a pattern! We encountered many beautiful examples of substrate-independent patterns in chapter 2, including waves, memories and computations. We saw how they weren't merely more than their parts (emergent), but rather independent of their parts, taking on a life of their own. For example, we saw how a future simulated mind or computer-game character would have no way of knowing whether it ran on Windows, Mac OS, an Android phone or some other operating system, because it would be substrate-independent. Nor could it tell whether the logic gates of its computer were made of transistors, optical circuits or other hardware. Or what the fundamental laws of physics are—they could be anything as long as they allow the construction of universal computers.

In summary, I think that consciousness is a physical phenomenon that feels non-physical because it's like waves and computations: it has properties independent of its specific physical substrate. This follows logically from the consciousness-as-information idea. This leads to a radical idea that I really like: If consciousness is the way that information feels when it's processed in certain ways, then it must be substrate-independent; it's only the structure of the information processing that matters, not the structure of the matter doing the information processing. In other words, consciousness is substrate-independent twice over!

As we've seen, physics describes patterns in spacetime that correspond to particles moving around. If the particle arrangements obey certain principles, they give rise to emergent phenomena that are pretty independent of the particle substrate, and have a totally different feel to them. A great example of this is information processing, in computronium. But we've now taken this idea to another level: *If the information processing itself obeys certain principles, it can give rise to the higher-level emergent phenomenon that we call consciousness.* This places your conscious experience not one but two levels up from the matter. No wonder your mind feels non-physical!

This raises a question: What are these principles that information processing needs to obey to be conscious? I don't pretend to know what conditions are *sufficient* to guarantee consciousness, but here are four *necessary* conditions that I'd bet on and have explored in my research:

Principle	Definition
Information principle	A conscious system has substantial information-storage capacity.
Dynamics principle	A conscious system has substantial information-processing capacity.
Independence principle	A conscious system has substantial independence from the rest of the world.
Integration principle	A conscious system cannot consist of nearly independent parts.

As I said, I think that consciousness is the way information feels when being processed in certain ways. This means that to be conscious, a system needs to be able to store and process information, implying the first two principles. Note that the memory doesn't need to last long: I recommend watching this touching video of Clive Wearing, who appears perfectly conscious even though his memories last less than a minute.[21] I think that a conscious system also needs to be fairly independent from the rest of the world, because otherwise it wouldn't subjectively feel that it had any independent existence whatsoever. Finally, I think that the conscious system needs to be integrated into a unified whole, as Giulio Tononi argued, because if it consisted of two independent parts, then they would feel like two separate conscious entities, rather than one. The first three principles imply *autonomy:* that the system is able to retain and process information without much outside interference, hence determining its own future. All four principles together mean that a system is autonomous but its parts aren't.

If these four principles are correct, then we have our work cut out for us: we need to look for mathematically rigorous theories that embody them and test them experimentally. We also need to determine whether additional principles are needed. Regardless of whether IIT is correct or not, researchers should try to develop competing theories and test all available theories with ever better experiments.

Controversies of Consciousness

We've already discussed the perennial controversy about whether consciousness research is unscientific nonsense and a pointless waste of time. In addition, there are recent controversies at the cutting edge of consciousness research—let's explore the ones that I find most enlightening.

Giulio Tononi's IIT has lately drawn not merely praise but also criticism, some of which has been scathing. Scott Aaronson recently had this to say on his blog: "In my opinion, the fact that Integrated Information Theory is wrong—demonstrably wrong, for reasons that go to its core—puts it in something like the top 2% of all mathematical theories of consciousness ever proposed. Almost all competing theories of consciousness, it seems to me, have been so vague, fluffy and malleable that they can only aspire to wrongness."[22] To the credit of both Scott and Giulio, they never came to blows when I watched them debate IIT at a recent New York University workshop, and they politely listened to each other's arguments. Aaronson showed that certain simple networks of logic gates had extremely high integrated information (Φ) and argued that since they clearly weren't conscious, IIT was wrong. Giulio countered that if they were built, they *would* be conscious, and that Scott's assumption to the contrary was anthropocentrically biased, much as if a slaughterhouse owner claimed that animals couldn't be conscious just because they couldn't talk and were very different from humans. My analysis, with which they both agreed, was that they were at odds about whether integration was merely a *necessary* condition for consciousness (which Scott was OK with) or also a *sufficient* condition (which Giulio claimed). The latter is clearly a stronger and more contentious claim, which I hope we can soon test experimentally.[23]

Another controversial IIT claim is that today's computer architectures can't be conscious, because the way their logic gates connect gives very low integration.[24] In other words, if you upload yourself into a future high-powered robot that accurately simulates every single one of your neurons and synapses, then even if this digital clone looks, talks and acts indistinguishably from you, Giulio claims that it will be an unconscious zombie without subjective experience—which would be disappointing if you uploaded yourself in a quest

for subjective immortality.[*6] This claim has been challenged by both David Chalmers and AI professor Murray Shanahan by imagining what would happen if you instead gradually replaced the neural circuits in your brain by hypothetical digital hardware perfectly simulating them.[25] Although your *behavior* would be unaffected by the replacement since the simulation is by assumption perfect, your *experience* would change from conscious initially to unconscious at the end, according to Giulio. But how would it feel in between, as ever more got replaced? When the parts of your brain responsible for your conscious experience of the upper half of your visual field were replaced, would you notice that part of your visual scenery was suddenly missing, but that you mysteriously knew what was there nonetheless, as reported by patients with "blindsight"?[26] This would be deeply troubling, because if you can consciously experience any difference, then you can also tell your friends about it when asked—yet by assumption, your behavior can't change. The only logical possibility compatible with the assumptions is that at exactly the same instance that any one thing disappears from your consciousness, your mind is mysteriously altered so as either to make you lie and deny that your experience changed, or to forget that things had been different.

On the other hand, Murray Shanahan admits that the same gradual-replacement critique can be leveled at *any* theory claiming that you can act conscious without being conscious, so you might be tempted to conclude that acting and being conscious are one and the same, and that externally observable behavior is therefore all that matters. But then you'd have fallen into the trap of predicting that you're unconscious while dreaming, even though you know better.

A third IIT controversy is whether a conscious entity can be made of parts that are separately conscious. For example, can society as a whole gain consciousness without the people in it losing theirs? Can a conscious brain have parts that are also conscious on their own? The prediction from IIT is a firm "no," but not everyone is convinced. For example, some patients with lesions severely reducing communication between the two halves of their brain experience "alien hand syndrome," where their right brain makes their left hand do things that the patients claim they aren't causing or understanding—sometimes to the point that they use their other hand to restrain their "alien" hand. How can we be so sure that there aren't two separate consciousnesses in their head, one in the right hemisphere that's unable to speak and another in the

left hemisphere that's doing all the talking and claiming to speak for both of them? Imagine using future technology to build a direct communication link between two human brains, and gradually increasing the capacity of this link until communication is as efficient between the brains as it is within them. Would there come a moment when the two individual consciousnesses suddenly disappear and get replaced by a single unified one as IIT predicts, or would the transition be gradual so that the individual consciousnesses coexisted in some form even as a joint experience began to emerge?

Another fascinating controversy is whether experiments underestimate how much we're conscious of. We saw earlier that although we *feel* we're visually conscious of vast amounts of information involving colors, shapes, objects and seemingly everything that's in front of us, experiments have shown that we can only remember and report a dismally small fraction of this.[27] Some researchers have tried to resolve this discrepancy by asking whether we may sometimes have "consciousness without access," that is, subjective experience of things that are too complex to fit into our working memory for later use.[28] For example, when you experience *inattentional blindness* by being too distracted to notice an object in plain sight, this doesn't imply that you had no conscious visual experience of it, merely that it wasn't stored in your working memory.[29] Should it count as forgetfulness rather than blindness? Other researchers reject this idea that people can't be trusted about what they say they experienced, and warn of its implications. Murray Shanahan imagines a clinical trial where patients report complete pain relief thanks to a new wonder drug, which nonetheless gets rejected by a government panel: "The patients only think they are not in pain. Thanks to neuroscience, we know better."[30] On the other hand, there have been cases where patients who accidentally awoke during surgery were given a drug to make them forget the ordeal. Should we trust their subsequent report that they experienced no pain?[31]

How Might AI Consciousness Feel?

If some future AI system is conscious, then what will it subjectively experience? This is the essence of the "even harder problem" of consciousness, and forces us up to the second level of difficulty depicted in figure 8.1. Not only do we currently lack a theory that answers this question, but we're not even sure whether it's logically possible to fully answer it. After all, what could a satisfactory answer sound like? How would you explain to a person born blind what the color red looks like?

Fortunately, our current inability to give a complete answer doesn't prevent us from giving partial answers. Intelligent aliens studying the human sensory system would probably infer that colors are qualia that feel associated with each point on a two-dimensional surface (our visual field), while sounds don't feel as spatially localized, and pains are qualia that feel associated with different parts of our body. From discovering that our retinas have three types of light-sensitive cone cells, they could infer that we experience three primary colors and that all other color qualia result from combining them. By measuring how long it takes neurons to transmit information across the brain, they could conclude that we experience no more than about ten conscious thoughts or perceptions per second, and that when we watch movies on our TV at twenty-four frames per second, we experience this not as a sequence of still images, but as continuous motion. From measuring how fast adrenaline is released into our bloodstream and how long it remains before being broken down, they could predict that we feel bursts of anger starting within seconds and lasting for minutes.

Applying similar physics-based arguments, we can make some educated guesses about certain aspects of how an artificial consciousness may feel. First of all, the space of possible AI experiences is *huge* compared to what we humans can experience. We have one class of qualia for each of our senses, but AIs can have vastly more types of sensors and internal representations of information, so we must avoid the pitfall of assuming that being an AI necessarily feels similar to being a person.

Second, a brain-sized artificial consciousness could have millions of times more experiences than us per second, since electromagnetic signals travel at the speed of light—millions of times faster than neuron signals. However, the larger

the AI, the slower its global thoughts must be to allow information time to flow between all its parts, as we saw in chapter 4. We'd therefore expect an Earth-sized "Gaia" AI to have only about ten conscious experiences per second, like a human, and a galaxy-sized AI could have only one global thought every 100,000 years or so—so no more than about a hundred experiences during the entire history of our Universe thus far! This would give large AIs a seemingly irresistible incentive to delegate computations to the smallest subsystems capable of handling them, to speed things up, much like our conscious mind has delegated the blink reflex to a small, fast and unconscious subsystem. Although we saw above that the conscious information processing in our brains appears to be merely the tip of an otherwise unconscious iceberg, we should expect the situation to be even more extreme for large future AIs: if they have a single consciousness, then it's likely to be unaware of almost all the information processing taking place within it. Moreover, although the conscious experiences that it enjoys may be extremely complex, they're also snail-paced compared to the rapid activities of its smaller parts.

This really brings to a head the aforementioned controversy about whether parts of a conscious entity can be conscious too. IIT predicts not, which means that if a future astronomically large AI is conscious, then almost all its information processing is unconscious. This would mean that if a civilization of smaller AIs improves its communication abilities to the point that a single conscious hive mind emerges, their much faster individual consciousnesses are suddenly extinguished. If the IIT prediction is wrong, on the other hand, the hive mind can coexist with the panoply of smaller conscious minds. Indeed, one could even imagine a nested hierarchy of consciousnesses at all levels from microscopic to cosmic.

As we saw above, the unconscious information processing in our human brains appears linked to the effortless, fast and automatic way of thinking that psychologists call "System 1."[32] For example, your System 1 might inform your consciousness that its highly complex analysis of visual input data has determined that your best friend has arrived, without giving you any idea how the computation took place. If this link between systems and consciousness proves to be valid, then it will be tempting to generalize this terminology to AIs, denoting all rapid routine tasks delegated to unconscious subunits as the AI's System 1. The effortful, slow and controlled global thinking of the AI would, if conscious, be the AI's System 2. We humans also have conscious experiences involving what I'll term "System 0": raw passive perception that takes place

even when you sit without moving or thinking and merely observe the world around you. Systems 0, 1 and 2 seem progressively more complex, so it's striking that only the middle one appears unconscious. IIT explains this by saying that raw sensory information in System 0 is stored in grid-like brain structures with very high integration, while System 2 has high integration because of feedback loops, where all the information you're aware of right now can affect your future brain states. On the other hand, it was precisely the conscious-grid prediction that triggered Scott Aaronson's aforementioned IIT-critique. In summary, if a theory solving the pretty hard problem of consciousness can one day pass a rigorous battery of experimental tests so that we start taking its predictions seriously, then it will also greatly narrow down the options for the even harder problem of what future conscious AIs may experience.

Some aspects of our subjective experience clearly trace back to our evolutionary origins, for example our emotional desires related to self-preservation (eating, drinking, avoiding getting killed) and reproduction. This means that it should be possible to create AI that never experiences qualia such as hunger, thirst, fear or sexual desire. As we saw in the last chapter, if a highly intelligent AI is programmed to have virtually any sufficiently ambitious goal, it's likely to strive for self-preservation in order to be able to accomplish that goal. If they're part of a society of AIs, however, they might lack our strong human fear of death: as long as they've backed themselves up, all they stand to lose are the memories they've accumulated since their most recent backup, as long as they're confident that their backed-up software will be used. In addition, the ability to readily copy information and software between AIs would probably reduce the strong sense of individuality that's so characteristic of our human consciousness: there would be less of a distinction between you and me if we could easily share and copy all our memories and abilities, so a group of nearby AIs may feel more like a single organism with a hive mind.

Would an artificial consciousness feel that it had free will? Note that, although philosophers have spent millennia quibbling about whether *we* have free will without reaching consensus even on how to define the question,[33] I'm asking a different question, which is arguably easier to tackle. Let me try to persuade you that the answer is simply "Yes, any conscious decision maker will subjectively *feel* that it has free will, regardless of whether it's biological or artificial." Decisions fall on a spectrum between two extremes:

1. You know exactly why you made that particular choice.
2. You have no idea why you made that particular choice—it felt like you chose randomly on a whim.

Free-will discussions usually center around a struggle to reconcile our goal-oriented decision-making behavior with the laws of physics: if you're choosing between the following two explanations for what you did, then which one is correct: *"I asked her on a date because I really liked her"* or *"My particles made me do it by moving according to the laws of physics"*? But we saw in the last chapter that *both* are correct: what feels like goal-oriented behavior can emerge from goal-less deterministic laws of physics. More specifically, when a system (brain or AI) makes a decision of type 1, it computes what to decide using some deterministic algorithm, and the reason it feels like it decided is that it in fact did decide when computing what to do. Moreover, as emphasized by Seth Lloyd,[34] there's a famous computer-science theorem saying that for almost all computations, there's no faster way of determining their outcome than actually running them. This means that it's typically impossible for you to figure out what you'll decide to do in a second in less than a second, which helps reinforce your experience of having free will. In contrast, when a system (brain or AI) makes a decision of type 2, it simply programs its mind to base its decision on the output of some subsystem that acts as a random number generator. In brains and computers, effectively random numbers are easily generated by amplifying noise. Regardless of where on the spectrum from 1 to 2 a decision falls, both biological and artificial consciousnesses therefore feel that they have free will: they feel that it is really they who decide and they can't predict with certainty what the decision will be until they've finished thinking it through.

Some people tell me that they find causality degrading, that it makes their thought processes meaningless and that it renders them "mere" machines. I find such negativity absurd and unwarranted. First of all, there's nothing "mere" about human brains, which, as far as I'm concerned, are the most amazingly sophisticated physical objects in our known Universe. Second, what alternative would they prefer? Don't they want it to be their own thought processes (the computations performed by their brains) that make their decisions? Their subjective experience of free will is simply how their computations feel from inside: they don't know the outcome of a computation until they've finished it. That's what it means to say that the computation *is* the decision.

Meaning

Let's end by returning to the starting point of this book: How do we want the future of life to be? We saw in the previous chapter how diverse cultures around the globe all seek a future teeming with positive experiences, but that fascinatingly thorny controversies arise when seeking consensus on what should count as positive and how to make trade-offs between what's good for different life forms. But let's not let those controversies distract us from the elephant in the room: there can be no positive experiences if there are no experiences at all, that is, if there's no consciousness. In other words, without consciousness, there can be no happiness, goodness, beauty, meaning or purpose—just an astronomical waste of space. This implies that when people ask about the meaning of life as if it were the job of our cosmos to give meaning to our existence, they're getting it backward: *It's not our Universe giving meaning to conscious beings, but conscious beings giving meaning to our Universe.* So the very first goal on our wish list for the future should be retaining (and hopefully expanding) biological and/or artificial consciousness in our cosmos, rather than driving it extinct.

If we succeed in this endeavor, then how will we humans feel about coexisting with ever smarter machines? Does the seemingly inexorable rise of artificial intelligence bother you and if so, why? In chapter 3, we saw how it should be relatively easy for AI-powered technology to satisfy our basic needs such as security and income as long as the political will to do so exists. However, perhaps you're concerned that being well fed, clad, housed and entertained isn't enough. If we're guaranteed that AI will take care of all our practical needs and desires, might we nonetheless end up feeling that we lack meaning and purpose in our lives, like well-kept zoo animals?

Traditionally, we humans have often founded our self-worth on the idea of *human exceptionalism:* the conviction that we're the smartest entities on the planet and therefore unique and superior. The rise of AI will force us to abandon this and become more humble. But perhaps that's something we should do anyway: after all, clinging to hubristic notions of superiority over others (individuals, ethnic groups, species and so on) has caused awful problems in the past, and may be an idea ready for retirement. Indeed, human exceptionalism

hasn’t only caused grief in the past, but it also appears unnecessary for human flourishing: if we discover a peaceful extraterrestrial civilization far more advanced than us in science, art and everything else we care about, this presumably wouldn’t prevent people from continuing to experience meaning and purpose in their lives. We could retain our families, friends and broader communities, and all activities that give us meaning and purpose, hopefully having lost nothing but arrogance.

As we plan our future, let’s consider the meaning not only of our own lives, but also of our Universe itself. Here two of my favorite physicists, Steven Weinberg and Freeman Dyson, represent diametrically opposite views. Weinberg, who won the Nobel Prize for foundational work on the standard model of particle physics, famously said, “The more the universe seems comprehensible, the more it also seems pointless.”[35] Dyson, on the other hand, is much more optimistic, as we saw in chapter 6: although he agrees that our Universe *was* pointless, he believes that life is now filling it with ever more meaning, with the best yet to come if life succeeds in spreading throughout the cosmos. He ended his seminal 1979 paper thus: “Is Weinberg’s universe or mine closer to the truth? One day, before long, we shall know.”[36] If our Universe goes back to being permanently unconscious because we drive Earth life extinct or because we let unconscious zombie AI take over our Universe, then Weinberg will be vindicated in spades.

From this perspective, we see that although we’ve focused on the future of intelligence in this book, the future of consciousness is even more important, since that’s what enables meaning. Philosophers like to go Latin on this distinction, by contrasting *sapience* (the ability to think intelligently) with *sentience* (the ability to subjectively experience qualia). We humans have built our identity on being *Homo sapiens*, the smartest entities around. As we prepare to be humbled by ever smarter machines, I suggest that we rebrand ourselves as *Homo sentiens*!

THE BOTTOM LINE:

- There's no undisputed definition of "consciousness." I use the broad and non-anthropocentric definition *consciousness = subjective experience.*
- Whether AIs are conscious in that sense is what matters for the thorniest ethical and philosophical problems posed by the rise of AI: Can AIs suffer? Should they have rights? Is uploading a subjective suicide? Could a future cosmos teeming with AIs be the ultimate zombie apocalypse?
- The problem of understanding intelligence shouldn't be conflated with three separate problems of consciousness: the "pretty hard problem" of predicting which physical systems are conscious, the "even harder problem" of predicting qualia, and the "*really* hard problem" of why anything at all is conscious.
- The "pretty hard problem" of consciousness is scientific, since a theory that predicts which of your brain processes are conscious is experimentally testable and falsifiable, while it's currently unclear how science could fully resolve the two harder problems.
- Neuroscience experiments suggest that many behaviors and brain regions are unconscious, with much of our conscious experience representing an after-the-fact summary of vastly larger amounts of unconscious information.
- Generalizing consciousness predictions from brains to machines requires a theory. Consciousness appears to require not a particular kind of particle or field, but a particular kind of information processing that's fairly autonomous and integrated, so that the whole system is rather autonomous but its parts aren't.
- Consciousness might feel so non-physical because it's doubly substrate-independent: if consciousness is the way information feels when being processed in certain complex ways, then it's merely the structure of the information processing that matters, not the structure of the matter doing the information processing.
- If artificial consciousness is possible, then the space of possible AI experiences is likely to be huge compared to what we humans can experience, spanning a vast spectrum of qualia and timescales—all sharing a feeling of having free will.
- Since there can be no meaning without consciousness, it's not our Universe giving meaning to conscious beings, but conscious beings giving meaning to our Universe.
- This suggests that as we humans prepare to be humbled by ever smarter machines, we take comfort mainly in being *Homo sentiens*, not *Homo sapiens.*

---

*1 An alternative viewpoint is *substance dualism*—that living entities differ from inanimate ones because

they contain some non-physical substance such as an “anima,” “élan vital” or “soul.” Support for substance dualism among scientists has gradually dwindled. To understand why, consider that your body is made of about 10^{29} quarks and electrons, which, as far as we can tell, move according to simple physical laws. Imagine a future technology able to track all your particles: if they were found to obey the laws of physics exactly, then your purported soul is having no effect on your particles, so your conscious mind and its ability to control your movements would have nothing to do with a soul. If your particles were instead found not to obey the known laws of physics because they were being pushed around by your soul, then the new entity causing these forces would by definition be a physical one that we can study just like we’ve studied new fields and new particles in the past.

*2 I use the word “qualia” according to the dictionary definition, to mean individual instances of subjective experience—that is, to mean the subjective experience itself, not any purported substance causing the experience. Beware that some people use the word differently.

*3 I’d originally called the RHP the “very hard problem,” but after I showed this chapter to David Chalmers, he emailed me the clever suggestion of switching to the *“really hard problem,”* to match what he really meant: “since the first two problems (at least put this way) aren’t really part of the hard problem as I conceived of it whereas the third problem is, you could perhaps use ‘really hard’ instead of ‘very hard’ for the third one to match my usage.”

*4 If our physical reality is entirely mathematical (information-based, loosely speaking), as I explored in my book *Our Mathematical Universe*, then no aspect of reality—not even consciousness—lies beyond the purview of science. Indeed, the really hard problem of consciousness is, from that perspective, the exact same problem as that of understanding how something mathematical can feel physical: if part of a mathematical structure is conscious, then it will experience the rest as its external physical world.

*5 Although I’ve earlier used “perceptronium” as a synonym for sentronium, that name suggests too narrow a definition, since percepts are merely those subjective experiences that we perceive based on sensory input—excluding, for example, dreams and internally generated thoughts.

*6 There’s potential tension between this claim and the idea that consciousness is substrate-independent, since even though the information processing may be different at the lowest level, it’s by definition identical at the higher levels where it determines behavior.