NEUROSCIENCE

What Is Consciousness?

Scientists are beginning to unravel a mystery that has long vexed philosophers

By Christof Koch

CONSCIOUSNESS IS EVERYTHING YOU experience. It is the tune stuck in your head, the sweetness of chocolate mousse, the throbbing pain of a toothache, the fierce love for your child and the bitter knowledge that eventually all feelings will end.

The origin and nature of these experiences, sometimes referred to as qualia, have been a mystery from the earliest days of antiquity right up to the present. Many modern analytic philosophers of mind, most prominently perhaps Daniel Dennett of Tufts University, find the existence of consciousness such an intolerable affront to what they believe should be a meaningless universe of matter and the void that they declare it to be an illusion. That is, they either deny that qualia exist or argue that they can never be meaningfully studied by science.

If that assertion was true, this essay would be very short. All I would need to explain is why you, I and most everybody else is so convinced that we have feelings at all. If I have a tooth abscess, however, a sophisticated argument to persuade me that my pain is delusional will not lessen its torment one iota. As I have very little sympathy for this desperate solution to the mind-body problem, I shall move on.

The majority of scholars accept consciousness as a given and seek to understand its relationship to the objective world described by science. More than a quarter of a century ago Francis Crick and I decided to set aside philosophical discussions on consciousness (which have engaged scholars since at least the time of Aristotle) and instead search for its physical footprints. What is it about a highly excitable piece of brain matter that gives rise to consciousness? Once we can understand that, we hope to get closer to solving the more fundamental problem.

We seek, in particular, the neuronal correlates of consciousness (NCC), defined as the minimal neuronal mechanisms jointly sufficient for any specific conscious experience. What must happen in your brain for you to experience a toothache, for example? Must some nerve cells vibrate at some magical frequency? Do some special "consciousness neurons" have to be activated? In which brain regions would these cells be located?

NEURONAL CORRELATES OF CONSCIOUSNESS

WHEN DEFINING THE NCC, the qualifier "minimal" is important. The brain as a whole can be considered an NCC, after all: it generates experience, day in and day out. But the seat of consciousness can be further ring-fenced. Take the spinal cord, a foot-and-a-half-long flexible tube of nervous tissue inside the backbone with about a billion nerve cells. If the spinal cord is completely severed by trauma to the neck region, victims are paralyzed in legs, arms and torso, unable to control their bowel and bladder, and without bodily sensations. Yet these tetraplegics continue to experience life in all its variety—they see, hear, smell, feel emotions and remember as much as before the incident that radically changed their life.

Or consider the cerebellum, the "little brain" underneath the back of the brain. One of the most ancient brain circuits in evolutionary terms, it is involved in motor control, posture and gait and in the fluid execution of complex sequences of motor movements. Playing the piano, typing, ice dancing or climbing a rock wall—all these activities involve the cerebellum. It has the brain's most glorious neurons, called Purkinje cells, which possess tendrils that spread like a sea fan coral and harbor complex electrical dynamics. It also has by far the most neurons, about 69 billion (most of which are the star-shaped cerebellar granule cells), four times more than in the rest of the brain combined.

What happens to consciousness if parts of the cerebellum are lost to a stroke or to the surgeon's knife? Very little! Cerebellar patients complain of several deficits, such as the loss of fluidity of piano playing or keyboard typing but never of losing any aspect of their consciousness. They hear, see and feel fine, retain a sense of self, recall past events and continue to project themselves into the future. Even being born without a cerebellum does not appreciably affect the conscious experience of the individual.

All of the vast cerebellar apparatus is irrelevant to subjective experience. Why? Important hints can be found within its circuitry, which is exceedingly uniform and parallel (just as batteries may be connected in parallel). The cerebellum is almost exclusively a feed-forward circuit: one set of neurons feeds the next, which in turn influences a third set. There are no complex feedback loops that reverberate with electrical activity passing back and forth. (Given the time needed for a conscious perception to develop, most theoreticians infer that it must involve feedback loops within the brain's cavernous circuitry.) More-

over, the cerebellum is functionally divided into hundreds or more independent computational modules. Each one operates in parallel, with distinct, nonoverlapping inputs and output, controlling movements of different motor or cognitive systems. They scarcely interact—another feature held indispensable for consciousness.

One important lesson from the spinal cord and the cerebellum is that the genie of consciousness does not just appear when any neural tissue is excited. More is needed. This additional factor is found in the gray matter making up the celebrated cerebral cortex, the outer surface of the brain. It is a laminated sheet of intricately interconnected nervous tissue, the size and width of a 14-inch pizza. Two of these sheets, highly folded, along with their hundreds of millions of wires—the white matter—are crammed into the skull. All available evidence implicates neocortical tissue in generating feelings.

We can narrow down the seat of consciousness even further. Take, for example, experiments in which different stimuli are presented to the right and the left eyes. Suppose a picture of Donald Trump is visible only to your left eye and one of Hillary Clinton only to your right eye. We might imagine that you would see some weird superposition of Trump and Clinton. In reality, you will see Trump for a few seconds, after which he will disappear and Clinton will appear, after which she will go away and Trump will reappear. The two images will alternate in a never-ending dance because of what neuroscientists call binocular rivalry. Because your brain is getting an ambiguous input, it cannot decide: Is it Trump, or is it Clinton?

If, at the same time, you are lying inside a magnetic scanner that registers brain activity, experimenters will find that a broad set of cortical regions, collectively known as the posterior hot zone, is active. These are the parietal, occipital and temporal regions in the posterior part of cortex [*see box on opposite page*] that play the most significant role in tracking what we see. Curiously, the primary visual cortex that receives and passes on the information streaming up from the eyes does not signal what the subject sees. A similar hierarchy of labor appears to be true of sound and touch: primary auditory and primary somatosensory cortices do not directly contribute to the content of auditory or somatosensory experience. Instead it is the next stages of processing—in the posterior hot zone—that give rise to conscious perception, including the image of Trump or Clinton.

More illuminating are two clinical sources of causal evidence: electrical stimulation of cortical tissue and the study of patients following the loss of specific regions caused by injury or disease. Before removing a brain tumor or the locus of a patient's epileptic seizures, for example, neurosurgeons map the functions of nearby cortical tissue by directly stimulating it with electrodes. Stimulating the posterior hot zone can trigger a diversity of distinct sensations and feelings. These could be flashes of light, geometric shapes, distortions of faces, auditory or visual hallucinations, a feeling of familiarity or unreality, the urge to move a specific limb, and so on. Stimulating the front of the cortex is a different matter: by and large, it elicits no direct experience.

A second source of insights are neurological patients from the first half of the 20th century. Surgeons sometimes had to excise a large belt of prefrontal cortex to remove tumors or to ameliorate epileptic seizures. What is remarkable is how unremarkable these patients appeared. The loss of a portion of the frontal lobe did have certain deleterious effects: the patients developed a lack of inhibition of inappropriate emotions or actions, motor deficits, or uncontrollable repetition of specific action or words. Following the operation, however, their personality and IQ improved, and they went on to live for many more years, with no evidence that the drastic removal of frontal tissue significantly affected their conscious experience. Conversely, removal of even small regions of the posterior cortex, where the hot zone resides, can lead to a loss of entire classes of conscious content: patients are unable to recognize faces or to see motion, color or space.

So it appears that the sights, sounds and other sensations of life as we experience it are generated by regions within the posterior cortex. As far as we can tell, almost all conscious experiences have their origin there. What is the crucial difference between these posterior regions and much of the prefrontal cortex, which does not directly contribute to subjective content? The truth is that we do not know. Even so—and excitingly—a recent finding indicates that neuroscientists may be getting closer.

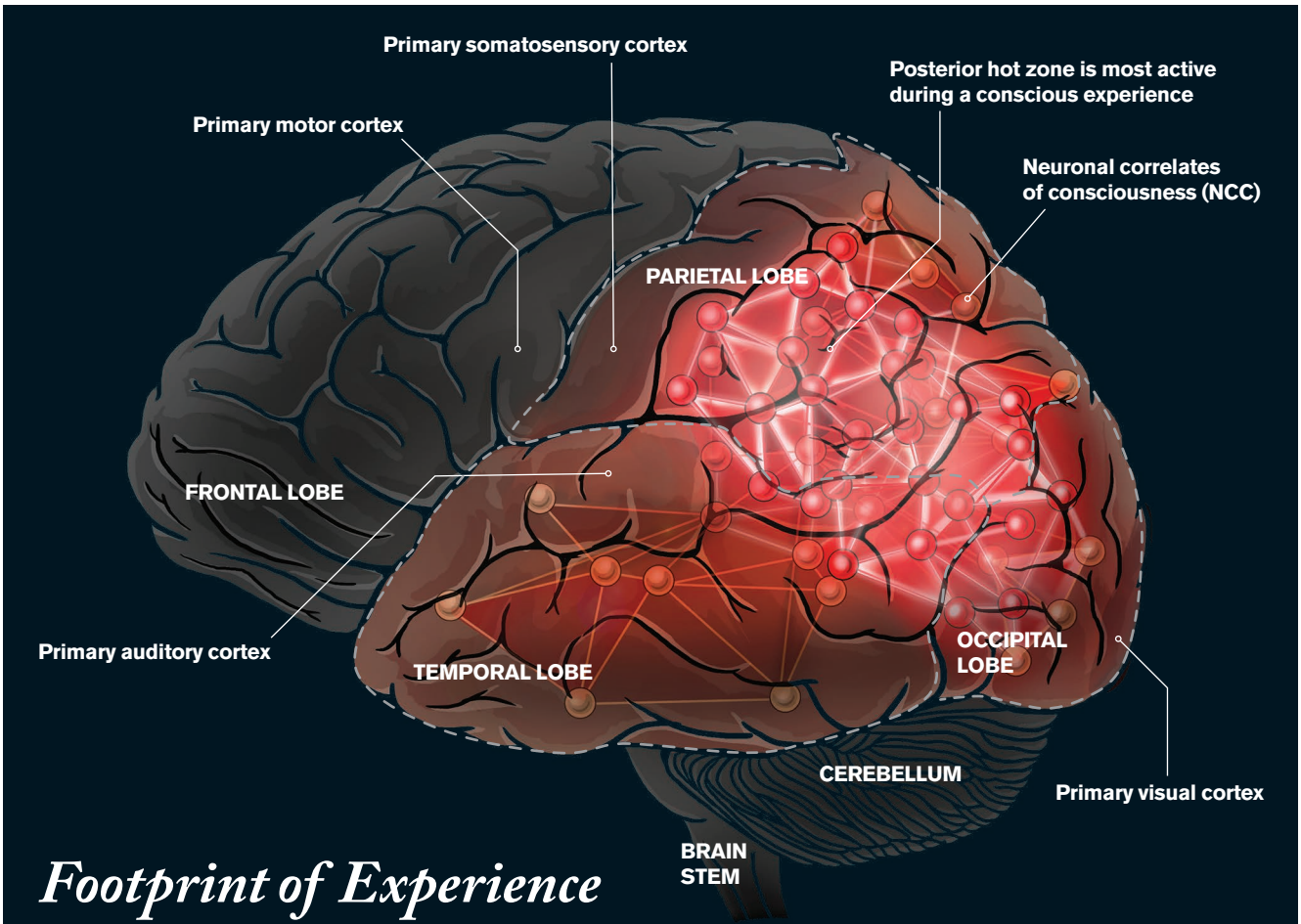
THE CONSCIOUSNESS METER

AN UNMET CLINICAL NEED exists for a device that reliably detects the presence or absence of consciousness in impaired or incapacitated individuals. During surgery, for example, patients are anesthetized to keep them immobile and their blood pressure stable and to eliminate pain and traumatic memories. Unfortunately, this goal is not always met: every year hundreds of patients have some awareness under anesthesia.

Another category of patients, who have severe brain injury because of accidents, infections or extreme intoxication, may live for years without being able to speak or respond to verbal requests. Establishing that they experience life is a grave challenge to the clinical arts. Think of an astronaut adrift in space, listening to mission control's attempts to contact him. His damaged radio does not relay his voice, and he appears lost to the world. This is the forlorn situation of patients whose damaged brain will not let them communicate to the world—an extreme form of solitary confinement.

In the early 2000s Giulio Tononi of the University of Wisconsin–Madison and Marcello Massimini, now at the University of Milan in Italy, pioneered a technique, called zap and zip, to probe whether someone is conscious or not. The scientists held a sheathed coil of wire against the scalp and "zapped" it—sent an intense pulse of magnetic energy into the skull—inducing a brief electric current in the neurons underneath. The perturbation, in turn, excited and inhibited the neurons' partner cells in connected regions, in a chain reverberating across the cortex, until the activity died out. A network of electroencephalogram (EEG) sensors, positioned outside the skull, recorded these electrical signals. As they unfolded over time, these traces, each corresponding to a specific location in the brain below the skull, yielded a movie.

These unfolding records neither sketched a stereotypical pattern, nor were they completely random. Remarkably, the more predictable these waxing and waning rhythms were, the more likely the brain was unconscious. The researchers quantified this intuition by



Footprint of Experience

Conscious awareness is closely associated with the cerebral cortex, an intricately folded and connected sheet of nervous tissue. Each experience corresponds to a specific set of neural activities, called the neuronal correlates of consciousness (NCC), in a posterior hot zone of the brain that consists of the parietal, occipital and temporal lobes of the cerebral cortex. Complexity of the neural excitations after a magnetic pulse yields a measure of the degree to which a person is conscious.

compressing the data in the movie with an algorithm commonly used to "zip" computer files. The zipping yielded an estimate of the complexity of the brain's response. Volunteers who were awake turned out have a "perturbational complexity index" of between 0.31 and 0.70, dropping to below 0.31 when deeply asleep or anesthetized. Massimini and Tononi tested this zap-and-zip measure on 48 patients who were brain-injured but responsive and awake, finding that in every case, the method confirmed the behavioral evidence for consciousness.

The team then applied zap and zip to 81 patients who were minimally conscious or in a vegetative state. For the former group, which showed some signs of nonreflexive behavior, the method correctly found 36 out of 38 patients to be conscious. It misdiagnosed two patients as unconscious. Of the 43 vegetative-state patients in which all bedside attempts to establish communication failed, 34 were labeled as unconscious, but nine were not. Their brains responded similarly to those of conscious controls—implying that they were conscious yet unable to communicate with their loved ones.

Ongoing studies seek to standardize and improve zap and zip for neurological patients and to extend it to psychiatric and pediatric patients. Sooner or later scientists will discover the specific set of neural mechanisms that give rise to any one experience. Although these findings will have important clinical implications and may give succor to families and friends, they will not answer some fundamental questions: Why these neurons and not those? Why this particular frequency and not that? Indeed, the abiding mystery is how and why any highly organized piece of active matter gives rise to conscious sensation. After all, the brain is like any other organ, subject to the same physical laws as the heart or the liver. What makes it different? What is it about the biophysics of a chunk of highly excitable brain matter that turns gray goo into the glorious surround sound and Technicolor that is the fabric of everyday experience?

Ultimately what we need is a satisfying scientific theory of consciousness that predicts under which conditions any particular physical system—whether it is a complex circuit of neurons or silicon transistors—has experiences. Furthermore, why does the quality of these experiences differ? Why does a clear blue sky feel so different from the screech of a badly tuned violin? Do these differences

Illustration by Mesa Schumacher

in sensation have a function, and if so, what is it? Such a theory will allow us to infer which systems will experience anything. Absent a theory with testable predictions, any speculation about machine consciousness is based solely on our intuition, which the history of science has shown is not a reliable guide.

Fierce debates have arisen around the two most popular theories of consciousness. One is the global neuronal workspace (GNW) by psychologist Bernard J. Baars and neuroscientists Stanislas Dehaene and Jean-Pierre Changeux. The theory begins with the observation that when you are conscious of something, many different parts of your brain have access to that information. If, on the other hand, you act unconsciously, that information is localized to the specific sensory motor system involved. For example, when you type fast, you do so automatically. Asked how you do it, you would not know: you have little conscious access to that information, which also happens to be localized to the brain circuits linking your eyes to rapid finger movements.

What is it about the biophysics of a chunk of brain matter that turns gray goo into the glorious surround sound and Technicolor that is the fabric of everyday experience?

TOWARD A FUNDAMENTAL THEORY

GNW ARGUES THAT CONSCIOUSNESS ARISES from a particular type of information processing—familiar from the early days of artificial intelligence, when specialized programs would access a small, shared repository of information. Whatever data were written onto this "blackboard" became available to a host of subsidiary processes: working memory, language, the planning module, and so on. According to GNW, consciousness emerges when incoming sensory information, inscribed onto such a blackboard, is broadcast globally to multiple cognitive systems—which process these data to speak, store or call up a memory or execute an action.

Because the blackboard has limited space, we can only be aware of a little information at any given instant. The network of neurons that broadcast these messages is hypothesized to be located in the frontal and parietal lobes. Once these sparse data are broadcast on this network and are globally available, the information becomes conscious. That is, the subject becomes aware of it. Whereas current machines do not yet rise to this level of cognitive sophistication, this is only a question of time. GNW posits that computers of the future will be conscious.

Integrated information theory (IIT), developed by Tononi and his collaborators, including me, has a very different starting point: experience itself. Each experience has certain essential properties. It is intrinsic, existing only for the subject as its "owner"; it is structured (a yellow cab braking while a brown dog crosses the street); and it is specific—distinct from any other conscious experience, such as a particular frame in a movie. Furthermore, it is unified and definite. When you sit on a park bench on a warm, sunny day, watching children play, the different parts of the experience—the breeze playing in your hair or the joy of hearing your toddler laugh—cannot be separated into parts without the experience ceasing to be what it is.

Tononi postulates that any complex and interconnected mechanism whose structure encodes a set of cause-and-effect relationships will have these properties—and so will have some level of consciousness. It will feel like something from the inside. But if, like the cerebellum, the mechanism lacks integration and complexity, it will not be aware of anything. As IIT states it, consciousness is intrinsic causal power associated with complex mechanisms such as the human brain.

IIT theory also derives, from the complexity of the underlying interconnected structure, a single nonnegative number Φ (pronounced "*fy*") that quantifies this consciousness. If Φ is zero, the system does not feel like anything to be itself. Conversely, the bigger this number, the more intrinsic causal power the system possesses and the more conscious it is. The brain, which has enormous and highly specific connectivity, possesses very high Φ, which implies a high level of consciousness. IIT explains a number of observations, such as why the cerebellum does not contribute to consciousness and why the zap-and-zip meter works. (The quantity the meter measures is a very crude approximation of Φ.)

IIT also predicts that a sophisticated simulation of a human brain running on a digital computer cannot be conscious—even if it can speak in a manner indistinguishable from a human being. Just as simulating the massive gravitational attraction of a black hole does not actually deform spacetime around the computer implementing the astrophysical code, programming for consciousness will never create a conscious computer. Consciousness cannot be computed: it must be built into the structure of the system.

Two challenges lie ahead. One is to use the increasingly refined tools at our disposal to observe and probe the vast coalitions of highly heterogeneous neurons making up the brain to further delineate the neuronal footprints of consciousness. This effort will take decades, given the byzantine complexity of the central nervous system. The other is to verify or falsify the two, currently dominant, theories. Or, perhaps, to construct a better theory out of fragments of these two that will satisfactorily explain the central puzzle of our existence: how a three-pound organ with the consistency of tofu exudes the feeling of life.

Christof Koch is chief scientist and president of the Allen Institute for Brain Science in Seattle. He serves on *Scientific American*'s board of advisers and has authored many books, including *Consciousness: Confessions of a Romantic Reductionist* (2012).