The Quest for Consciousness: 
A Neurobiological Approach
by Christof Koch
Roberts & Company Publishers,
2004
429 pages, $45

This book stems from a 16-year collaboration between Christof Koch, the Troendle Professor of Cognitive and Behavioral Biology and professor of computation and neural systems and Nobel Laureate Francis Crick, of the Salk Institute in La Jolla, to find a neurological explanation for consciousness—one of the last major unsolved problems of science. Most of the book deals with Koch’s research on vision and perception, but the lively, informal style and thorough glossary make it accessible for anyone who has ever thought about thinking. What follows is the final chapter, which has been written in the form of a fictitious interview.

Interviewer: Let’s start at the beginning. What is the overall strategy that you are pursuing in tackling this problem?

Christof Koch: First, I take consciousness seriously, as a brute fact that needs to be explained. The first-person perspective, feelings, qualia, awareness, phenomenal experiences—call it what you want—are real phenomena that arise out of certain privileged brain processes. They make up the landscape of conscious life: the deep red of a sunset over the Pacific Ocean, the fragrance of a rose, the searing anger that wells up at seeing an abused dog, the memory of the exploding space shuttle Challenger on live TV. Science’s ability to comprehend the universe will be limited unless and until it can explain how certain physical systems can be sufficient for such subjective states.

Second, I argue for putting aside, for now, the difficult problems that philosophers debate—in particular the question of why it is that it feels like something to see, hear, or to be me—and concentrate on a scientific exploration of the molecular and neuronal correlates of consciousness (NCC). The question I focus on is, What are the minimal neuronal mechanisms jointly sufficient for a specific conscious percept? Given the amazing technologies that brain scientists have at their disposal—engineering the mammalian genome, simultaneously recording from hundreds of neurons in a monkey, imaging the living human brain—the search for the neuronal correlates of consciousness, the NCC, is tractable, clearly defined, and will yield to a concerted scientific attack.

Do you mean to imply that discovering the NCC will solve the mystery of consciousness?

No, no, no! Ultimately, what is needed is a principled account explaining why and under what circumstances certain types of very complex biological entities have subjective experiences and why these experiences appear the way they do. The past two thousand years are littered with attempts to solve these mysteries, so they truly are hard problems.

Remember how much the elucidation of the double-helical structure of DNA revealed about molecular replication? The two complementary chains of sugar, phosphate, and amine bases, linked by weak hydrogen bonds, immediately suggested a mechanism whereby genetic information could be represented, copied, and passed on to the next generation. The architecture of the DNA molecule led to an understanding of heredity that was simply beyond the capabilities of the previous generations of chemists and biologists. By analogy, knowing where the neurons that mediate a specific conscious percept are located, where they project to and receive input from, their firing pattern, their developmental pedigree from birth to adulthood, and so on, might provide a similar breakthrough on the way to a complete theory of consciousness.

A fond dream.

Perhaps so, but there is no credible alternative to understanding consciousness by searching for the NCC. Experience has shown that logical argumentation and introspection, the preferred methods of scholars throughout all but the past two centuries, are simply not powerful enough to crack this problem. You can’t reason your way to an explanation of consciousness. Brains are too complicated, and are conditioned on too many random events and accidents of evolutionary history, for such armchair methods to successfully illuminate the truth. Instead, you have to find out the facts. How specific is the tapestry woven by axons among neurons? Does synchronized firing play a critical role in the genesis of consciousness? How crucial are the feedback pathways crisscrossing cortex and thalamus? Are there special neuronal cell types that underlie the NCC?

What, then, is the role of philosophers in your quest for a scientific theory of consciousness?

Historically, philosophy does not have an impressive track record of answering questions about the natural world in a decisive manner, whether it’s the origin and evolution of the
cosmos, the origin of life, the nature of the mind, or the nature-versus-nurture debate. This failure is rarely talked about in polite, academic company. Philosophers, however, excel at asking conceptual questions from a point of view that scientists don’t usually consider. Notions of the Hard versus the Easy Problem of consciousness, phenomenal versus access consciousness, the content of consciousness versus consciousness as such, the unity of consciousness, the causal conditions for consciousness to occur, and so on, are fascinating issues that scientists should ponder more often. So, listen to the questions posed by philosophers but don’t be distracted by their answers. A case in point is the philosopher’s zombie.

**Zombies? Cursed, dead people walking around with outstretched arms?**

Well, no. People like you and me but with no conscious feelings at all. David Chalmers and other philosophers use these soulless, fictitious creatures to argue that consciousness does not follow from the physical laws of the universe; that knowing about physics, biology, and psychology won’t help one iota in understanding how and why experience enters the universe. Something more is required.

This radical, imaginary zombie doesn’t strike me as a very useful concept; but there is a more modest and restricted version. Therefore, Francis and I co-opted this catchy term for the set of rapid, stereotyped sensory-motor behaviors that are insufficient, by themselves, for conscious sensations. The classic example is motor control. When you want to run along a trail, you “just do it.” Proprioceptive sensors, neurons, and the muscular-skeletal system take care of the rest, and you’re on your way. Try to introspect and you’ll be confronted with a blank wall. Consciousness has no access to the amazingly complex sequence of computations and actions that underlie such a seemingly simple behavior.

**So zombie behaviors are reflexes, only more complex?**

Yes. Think of them as cortical reflexes. Reaching for a glass of water by extending your arm and automatically opening the hand to grasp it constitutes a zombie action that requires visual input to control the arm and hand. You carry out thousands of these actions daily. You can “see” the glass, of course, but only because neural activity in a different system is responsible for the conscious percept.

**You imply that unconscious, zombie systems co-exist with conscious ones in normal, healthy folks.**

Exactly. A disconcertingly large fraction of your everyday behavior is zombie-like: You drive to work on autopilot, move your eyes, brush your teeth, tie your shoelaces, greet your colleagues in the hall, and perform all the other myriad chores that constitute daily life. Any sufficiently well-rehearsed activity, such as rock climbing, dancing, martial arts, or tennis is best performed without conscious, deliberate thought. Reflecting too much about any one action will interfere with its seamless execution.

**Why, then, is consciousness necessary at all? Why couldn’t I be a zombie?**

Well, I know of no logical reason why you couldn’t, although life would be pretty boring without any sensations (of course, you wouldn’t feel any ennui as a zombie). However, evolution took a different turn on this planet. Some simple creatures may be nothing but bundles of zombie agents. Thus, it might not feel like anything to be a snail or a roundworm.
However, you happen to be an organism with plenty of input sensors and output effectors, say, a mammal, devoting a zombie system to each and every possible input-output combination became too expensive. It would have taken up too much room in the skull. Instead, evolution chose a different path, evolving a powerful and flexible system whose primary responsibility is to deal with the unexpected and to plan for the future. The NCC represent selected aspects of the environment—the ones you are currently aware of—in a compact manner. This information is made accessible to the planning stages of the brain, with the help of some form of immediate memory.

In computer lingo, the current content of awareness corresponds to the state of cache memory on the CPU. As your stream of consciousness flitters from a visual perpect to a memory to a voice out there, the content of the cache fluctuates, too.

**As in the meaning of a word?**

No, not in any linguistic sense. The objects I feel, see, or hear out there in the world are not meaningless symbols but come with rich associations. The bluish tinge of a fine porcelain cup brings back childhood memories. I know I can grab the cup and pour tea into it. If it falls to the ground it will shatter. These associations don’t have to be made explicit. They are built up from countless sensory-motor interactions with the world over a lifetime of experiences. This elusive meaning corresponds to the sum total of all synaptic interactions of the neurons representing the porcelain cup with neurons expressing other concepts and memories. All the vast information is symbolized, in a shorthand way, by the qualia associated with the perect of the cup. That’s what you experience.

Leaving that aside for now, what is so important in this field, which has been plagued by hundreds of years of unsubstantiated speculations, is that our framework leads to tests for consciousness. Zombie agents operate in the here and now, so they have no need for short-term memory. You see an outstretched hand, so you reach out and shake it with your own hand. A zombie could not handle a delay between the sight of the hand and the motor action; it didn’t evolve to deal with that. The more powerful, albeit slower, consciousness system would have to take over.

These different behaviors can be shaped into a simple operationalized test for consciousness in animals, babies, or patients that can’t easily communicate their experiences. Force the organism to make a choice, such as inhibiting an instinctual behavior, following a delay of a few seconds. If the creature can do so without extensive learning, it must make use of a planning module that, at least in humans, is closely linked to consciousness. If the NCC underlying this action is destroyed (or rendered inoperable for some time) by some external means, the delayed response shouldn’t happen anymore.

---

**The Turing test**

was proposed in 1950 by mathematician Alan Turing as a way of testing the intelligence of a machine. If, after carrying out an extended online conversation on a variety of arbitrary topics, you can’t tell whether you’ve been interacting with a human or a machine, the machine should be considered intelligent.

---

Given the close evolutionary kinship among mammals, and the structural similarity of their brains, I assume that monkeys, dogs, and cats can be aware of what they see, hear, or smell.

**Wait. Are you saying that insects may be conscious?**

Many scholars believe that consciousness requires language and a representation of the self as a basis for introspection. While there is no doubt that humans can recursively think about themselves, this is just the latest elaboration of a more basic biological phenomenon that evolved a long time ago.

Consciousness can be associated with quite elemental feelings. You see purple or have pain. Why should these sensations require language or a highly developed notion of the self? Even severely autistic children or patients with massive self-delusions and depersonalization syndromes don’t lack basic perceptual awareness—the ability to see, hear, or smell the world.

The pre-linguistic origin of perceptual consciousness, the type of consciousness I study, raises the question of how far down the evolutionary ladder it extends. At what point in time did the Ur-NCC first appear? Given the close evolutionary kinship among mammals, and the structural similarity of their brains, I assume that monkeys, dogs, and cats can be aware of what they see, hear, or smell.
What about mice, the most popular mammal in biological and medical laboratories?

Given the comparative ease of manipulating the mouse genome, of inserting new genes or knocking out existing ones, applying the anti-zombie delay test to mice in some practical manner would give molecular neuroscientists a powerful model to study the basis of the NCC. My laboratory and others are developing such a mouse model of attention and awareness using classical Pavlovian conditioning.

Wait. Why did you say “awareness” instead of “consciousness”? Do they refer to different concepts?

No. It is more of a social convention. Consciousness—the C word—evokes powerful aversive reactions in some colleagues; so you’re better off with some other word in grant applications and journal submissions. “Awareness” usually slips under the radar.

Continuing with animal consciousness, why stop at mice or, indeed, at mammals? Why be a cortical chauvinist? Do we really know that the cerebral cortex and its satellites are necessary for perceptual consciousness? Why not squids? Or bees? Endowed with one million neurons, bees can perform complicated actions, including amazing feats of visual pattern matching. For all I know, a hundred thousand neurons may be sufficient to see, to smell, and to feel pain! Maybe even fruit flies are conscious, to a very limited extent. Today we just don’t know.

Sounds like unsubstantiated speculations to me.

For now, yes. But behavioral and physiological experiments bring these speculations into the realm of the empirical. And this is new. We were not in a position to think about such litmus tests until recently.

Could these tests be applied to machines to assess whether they are conscious?

I’m not only a member of the biology faculty at Caltech, but also a professor in the Division of Engineering and Applied Science, so I do think about artificial consciousness, based on an analogy to neurobiology. Any organism capable of behaviors that go beyond the instinctual and that has some way to express the meaning of symbols is a candidate for sentence.

The Internet taken as a whole is a tantalizing example of an emergent system with millions of computers acting as nodes in a distributed, but highly interconnected network. While there are file swapping programs that link large numbers of computers, or algorithms that solve mathematically intractable problems by distributing them over thousands of machines, these assemblies bear little relationship to the coalitions of neurons that excite and inhibit each other in the brain. There are no collective behaviors of the entire World Wide Web to speak of. I’ve never witnessed the spontaneous appearance of any purposeful, large-scale action not designed into the software. It doesn’t make any sense to speak of the conscious Web unless it displays such behaviors on its own—by directing electrical power allocation, controlling airline traffic, or manipulating financial markets in a manner unintended by its makers. With the emergence of autonomous computer viruses and worms, this may change in the future, though.

What about a robot endowed with reflex-like behaviors—to avoid running into obstacles, to prevent its battery from draining, to communicate with other robots, and so on—in addition to a general planning module. Could that be conscious?

Well, suppose the planner was powerful enough to represent the machine’s current sensory environment, including its own body and some of the information retrieved from its memory banks that is germane to the present situation, so that it would be capable of independent and purposeful behaviors. Assume, moreover, that your robot could learn to relate sensory events to positive and negative goal states so as to guide its behavior. A high ambient temperature, for example, might cause a drop in the machine’s supply voltage—something it would want to avoid at all costs. An elevated temperature wouldn’t be an abstract number anymore but would be intimately connected to the organism’s well-being. Such a robot might have some level of proto-consciousness.

That seems like quite a primitive notion of meaning.

Sure, but I doubt that at your birth you were conscious of much more than pain and pleasure. There are other sources of meaning, though. Imagine that the robot establishes sensory-motor representations by some unsupervised learning algorithm. It would stumble and fumble its way around the world and would learn, by trial and error, that its actions lead to predictable conse-
quences. At the same time, more abstract represen-
tations could be built up by comparing infor-
mation from two or more sensory modalities (e.g.,
that moving lips and particular staccato sound
patterns often go together). The more explicit
representations there are, the more meaningful
any one concept is.

To establish these meanings, it would be easiest
if the machine designers could replicate the
developmental phases of childhood for the robot.

Just like HAL, the paranoid computer in the movie
2001! But you haven’t answered my earlier question
yet. Would your delay test distinguish a truly conscious
machine from a fake that is just pretending to be conscious?

Just because this exercise distinguishes reflexive
systems from conscious ones in biological organ-
isms doesn’t imply that it will do the same for
machines.

It makes sense to grant at least some animal
species sentience due to their evolutionary,
behavioral, and structural similarity to humans,
based on an argument of the form “since I am
conscious, the more similar other organisms are to
me, the more likely they are to have feelings.”
This argument loses its power, though, in the face
of the radically different design, origin, and form
of machines.

Let’s leave this topic and look back to your earlier ideas
about the neuronal correlates of consciousness. What did
you and Francis propose?

In our first publication on the topic in 1990, we
put forth the idea that one form of consciousness
involves dynamic binding of neural activity across
multiple cortical areas.

Wait, wait. What’s binding?

Think of a red Ferrari zooming past. This
triggers nervous activity at myriad locations
throughout the brain, yet you see a single red
object in the shape of a car, moving in a certain
direction, and emitting a lot of noise. The
integrated percept has to combine the activity of
neurons that encode for the motion with neurons
that represent red and others that encode the shape
and the sounds. At the same time, you notice a
pedestrian with a dog walking past. This also has
to be expressed neuronally without confusing it
with the representation for the Ferrari.

At the time of our 1990 paper, two German
groups, led by Wolf Singer and Reinhardt Eckhorn,
respectively, had discovered that neurons in the cat
visual cortex, under certain conditions, would
synchronize their discharge patterns. Often, this
would occur periodically, giving rise to the famous
40 Hertz (Hz) oscillations. We argued that this
was one of the neuronal signatures of consciousness.

What does the evidence look like now?

The neuroscience community remains deeply
divided on the topic of oscillations and synchroni-

tization. A scientific journal will publish evidence
in favor of their functional relevance, while a
contribution in the following issue pooh-poohs the
entire concept. Unlike cold fusion, which has no
credible evidence in its favor, the basic existence
of neuronal oscillations in the 20 to 70 Hz frequency
range and synchronized discharges is accepted.
There is a great deal more, however, that remains
contentious. Our reading of the data is that
synchronized and oscillatory firing helps one
coalition—representing one percept—overcome
the others in the competition for dominance. Such
a mechanism might be particularly important
during attentional biasing. We no longer believe
that 40 Hz oscillations are necessary for conscious-
ness to occur.

This uncertainty is symptomatic of the inade-
quacy of existing tools to probe the neuronal
networks that underlie the mind. In a cortex of
billions of cells, state-of-the-art electrophysiologi-
cal techniques can listen to the pulses emanating
from a hundred neurons. That’s a dilution of one
out of one hundred million. What is needed is
the record of the simultaneous activity of ten thousand
or one hundred thousand brain cells.

So, if the NCC are based on a coalition of cells, their
existence could easily be missed among the din of those
billions of neurons chattering to each other.

Precisely. It’s like trying to learn something
meaningful about an upcoming presidential
election by recording the everyday conversations
of two or three randomly chosen people.

I see. Let’s move on to your next step.

This came in 1995 and pertained to the
function of consciousness, which we had ignored
up to that point. We hypothesized that a major
function of consciousness was to plan for the
future, allowing the organism to rapidly deal with
many contingencies. This, by itself, was not so
different from what other scholars had proposed.
We took this argument a step further and asked
about its neuroanatomical consequences. Because
the planning parts of the brain are located in the
frontal lobe, the NCC must have direct access to
these brain regions. It turns out that in the
monkey, none of the neurons within the primary
visual cortex, V1, at the back of the brain, send
their output to the front of the brain. We there-
fore concluded that V1 neurons are not sufficient
for visual perception, that visual consciousness
requires higher cortical regions.

That’s not to say that an intact V1 isn’t neces-
sary for seeing. Just as the neural activity in your
eyes does not correspond to visual perception—
only otherwise you would see a gray disk of
nothingness at the blind spot where the optic
nerve leaves the eye and no photoreceptors exist—
V1 activity is necessary but insufficient for sight.
V1 is probably not necessary for visual imagery or
for experiencing visual dreams.
I don’t see why you make such a big deal out of this. So what if the NCC aren’t in V1?

Well, if true—and the current evidence is quite encouraging—our hypothesis represents a modest but measurable step forward. This is emboldening because it demonstrates that, with the right approach, science can make progress in uncovering the material basis of consciousness. Our hypothesis also implies that not all cortical activity is expressed consciously.

So where, among the vast fields of the cortex, are the NCC?

Look within the “vision-for-perception” pathway if you’re concerned with visual consciousness. Coalitions of neurons in and around the inferior temporal cortex, supported by feedback activity from cells in the cingulate and frontal cortices, are essential. By way of this reverberatory feedback activity, the coalition can win out over its competitors. The echoes of this conflict can be picked up by EEG or functional brain imaging.

Ongoing electrophysiological explorations of these brain regions continue apace. A popular strategy exploits visual illusions in which the relationship between an image and its associated percept is not one-to-one. Although the input is continuously present, sometimes you see it one way and sometimes in another. Such bistable percepts—the Necker cube is the classical example—are used to track the footprints of consciousness among the different neuronal cell types in the forebrain.

Why invoke a loop from the sensory regions of the cortex to the more frontal ones?

As I just mentioned, this is one of the pivotal roles of consciousness in the life of an organism—to plan for multicontingency situations that can’t be dealt with by the nonconscious sensory-motor agents. It is probably the projections to and from the frontal lobes, responsible for planning, thought, and reasoning and the seat of the self, that create the powerful feeling that there is a homunculus inside my head, the true “me.” The little person—the original meaning of the term homunculus—is part of the front of the cortex observing the back. Or, in anatomical terms, the anterior cingulate, prefrontal, and premotor cortices are receiving a strong, driving synaptic input from the back of the cortex.

But who is, in turn, inside the homunculus’s head? Don’t you end up with an infinite loop?

Not if the homunculus is, itself, unconscious or has a reduced functional role compared to that of the conscious mind.

Can the homunculus freely initiate actions?

You must sharply distinguish the perception of will from the force of will. See, I can raise my hand and I certainly feel that I am willing this action. Nobody told me to and I didn’t even think about this until a few seconds ago. Perception of control, of authorship—the sense that I am in charge—is essential to my survival, enabling my brain to label these actions as mine (this perception of authorship will have its own NCC, of course). The neuropsychologist Daniel Wegner points out that the belief “I can initiate actions” is a form of optimism. It lets me accomplish things with confidence and exuberance that a pessimist might never attempt.

But was your raised hand completely determined by prior events or was it freely willed?

You mean, do the laws of physics leave room for a will that is free in the metaphysical sense? Everybody has opinions on this age-old problem, but there are no generally accepted answers. I do know of many instances of a dissociation between an individual’s action and her intentions. You can observe these slip-ups in your own life. When “you want” to climb above a ledge, for example, but your body doesn’t follow because it’s too scared. Or, when running in the mountains and your will slackens but your legs just keep on going. There are many extreme forms of dissociations between action and the experience of willing an action, including hypnosis, table turning, automatic writing, facilitated communications, spirit possession, deindividuation in crowds, and clinical dissociative identity disorders. But whether raising my hand was truly free, as free as Siegfried’s destruction of the world order of the gods in Wagner’s Der Ring des Nibelungen, I doubt it.
From your answer I gather, in any case, that you think your quest for the NCC can be divorced from the question of free will.

Yes. Whether or not free will exists, you still have to explain the puzzle of experience, of sensation.

What will be the consequences of discovering the NCC?

The most obvious ones will be of a practical nature, such as techniques to track the status of the NCC. Such a conscious-ometer will enable medical personnel to monitor the presence of consciousness in premature babies and young infants, in patients whose minds are afflicted with severe autism, or senile dementia, and in patients who are too injured to speak or even to signal. It will permit anesthesiologists to better practice their craft. Understanding the brain basis of consciousness will allow scientists to determine which species are sentient. Do all primates experience the sights and sounds of the world? All mammals? All multicellular organisms? This discovery should profoundly affect the animal rights debate.

How so?

Species without NCC can be thought of as bundles of stereotyped sensory-motor loops, without subjective experience, zombies. Such organisms could be accorded less protection than animals that do show NCCs under some conditions.

So, you would not want to experiment with animals that can feel pain?

In the ideal world, no. However, one of my daughters died 8 weeks after birth from sudden infant death syndrome; my father wasted away over a period of twelve years from Parkinson's disease compounded at the end by Alzheimer's disease; and a good friend killed herself in the throes of a florid episode of schizophrenia. Eliminating these and other neuronal maladies afflicting humanity requires animal experimentation—carried out with care and compassion and, whenever possible, with the animal's cooperation (as in the vast bulk of the monkey research described in this book).

What about implications for ethics and religion?

What matters from a metaphysical point of view is whether neuroscience can successfully move beyond correlation to causation. Science seeks a causal chain of events that leads from neural activity to subjective percept; a theory that accounts for what organisms under what conditions generate subjective feelings, what purpose these serve, and how they come about.

If such a theory can be formulated—a big if—without resorting to new ontological entities that can't be objectively defined and measured, then the scientific endeavor, dating back to the Renaissance, will have risen to its last great challenge. Humanity will have a closed-form, quantitative account of how mind arises out of matter. This is bound to have significant consequences for ethics, including a new conception of humans that might radically contradict the traditional images that men and women have made of themselves throughout the ages and cultures.

Not everybody will be enthralled by this. Many will argue that this success marks the nadir of science's relentless, dehumanizing drive to deprive the universe of meaning and significance.

But why? Why should knowledge lessen my appreciation of the world around me? I am in awe that everything I see, smell, taste, or touch is made out of 92 elements, including you, me, this book, the air we breathe, the earth we stand on, the stars in the sky. And these elements can be arranged in a periodic kingdom. This, in turn, rests on an even more fundamental triad of protons, neutrons, and electrons. What secret form of cabalistic knowledge provides greater satisfaction? And none of this intellectual understanding lessens my love of life and the people, dogs, nature, books, and music around me by one bit.

What about religion? Most people on the planet believe in some sort of immortal soul that lives on after the body has died. What do you have to say to them?

Well, many of these beliefs can't be reconciled with our current scientific world view. What is clear is that every conscious act or intention has some physical correlate. With the end of life, consciousness ceases, for without brain, there is no mind. Still, these irrevocable facts do not exclude some beliefs about the soul, resurrection, and God.

Now that your five-year-ordeal of writing this book is over and your children have left for college, what are you going to do?

As Maurice Herzog famously pealed at the end of Annapurna, his account of the first ascent of the eponymous Himalayan mountain, “There are other Annapurnas in the lives of men.”