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.

CONSCIOUSNESS

Proust among the Machines

Within our lifetimes, computers could approach human-level intelligence. But will they be able to consciously experience the world?

By Christof Koch

A FUTURE WHERE THE THINKING CAPABILITIES OF COMPUTERS APPROACH OUR OWN IS quickly coming into view. We feel ever more powerful machine-learning (ML) algorithms breathing down our necks. Rapid progress in coming decades will bring about machines with human-level intelligence capable of speech and reasoning, with a myriad of contributions to economics, politics and, inevitably, warcraft. The birth of true artificial intelligence will profoundly affect humankind's future, including whether it has one.

The following quotes provide a case in point:

"From the time the last great artificial intelligence breakthrough was reached in the late 1940s, scientists around the world have looked for ways of harnessing this 'artificial intelligence' to improve technology beyond what even the most sophisticated of today's artificial intelligence programs can achieve."

"Even now, research is ongoing to better understand what the new AI programs will be able to do, while remaining within the bounds of today's intelligence. Most AI programs currently programmed have been limited primarily to making simple decisions or performing simple operations on relatively small amounts of data."

These two paragraphs were written by GPT-2, a language bot I tried last summer. Developed by OpenAI, a San Francisco–based institute that promotes beneficial AI, GPT-2 is an ML algorithm with a seemingly idiotic task: presented with some arbitrary starter text, it must predict the next word. The network isn't taught to "understand" prose in any human sense. Instead, during its training phase, it adjusts the internal connections in its simulated neural networks to best anticipate the next word, the word after that, and so on. Trained on eight million Web pages, its innards contain more than a billion connections that emulate synapses, the connecting points between neurons. When I entered the first few sentences of the article you are reading, the algorithm spewed out two paragraphs that sounded like a freshman's effort to recall the gist of an introductory lecture on machine learning during which she was daydreaming. The output contains all the right words and phrases—not bad, really! Primed with the same text a second time, the algorithm comes up with something different.

The offspring of such bots will unleash a tidal wave of "deep-

Illustration by Gérard Dubois

fake" product reviews and news stories that will add to the miasma of the Internet. They will become just one more example of programs that do things hitherto thought to be uniquely human—playing the real-time strategy game StarCraft, translating text, making personal recommendations for books and movies, recognizing people in images and videos.

It will take many further advances in machine learning before an algorithm can write a masterpiece as coherent as Marcel Proust's *In Search of Lost Time,* but the code is on the wall. Recall that all early attempts at computer game playing, translation and speech were clumsy and easy to belittle because they so obviously lacked skill and polish. But with the invention of deep neural networks and the massive computational infrastructure of the tech industry, computers relentlessly improved until their outputs no longer appeared risible. As we have seen with Go, chess and poker, today's algorithms can best humans, and when they do, our initial laughter turns to consternation. Are we like Goethe's sorcerer's apprentice, having summoned helpful spirits that we now are unable to control?

ARTIFICIAL CONSCIOUSNESS?

ALTHOUGH EXPERTS DISAGREE over what exactly constitutes intelligence, natural or otherwise, most accept that, sooner or later, computers will achieve what is termed artificial general intelligence (AGI) in the lingo.

The focus on machine intelligence obscures quite different questions: Will it feel like anything to be an AGI? Can programmable computers ever be conscious?

By "consciousness" or "subjective feeling," I mean the quality inherent in any one experience—for instance, the delectable taste of Nutella, the sharp sting of an infected tooth, the slow passage of time when one is bored, or the sense of vitality and anxiety just before a competitive event. Channeling philosopher Thomas Nagel, we could say a system is conscious if there is something it is like to *be* that system.

Consider the embarrassing feeling of suddenly realizing that you have just committed a gaffe, that what you meant as a joke came across as an insult. Can computers ever experience such roiling emotions? When you are on the phone, waiting minute after minute, and a synthetic voice intones, "We are sorry to keep you waiting," does the software actually feel bad while keeping you in customer-service hell?

There is little doubt that our intelligence and our experiences are ineluctable consequences of the natural causal powers of our brain, rather than any supernatural ones. That premise has served science extremely well over the past few centuries as people explored the world. The three-pound, tofulike human brain is by far the most complex chunk of organized active matter in the known universe. But it has to obey the same physical laws as dogs, trees and stars. Nothing gets a free pass. We do not yet fully understand the brain's causal powers, but we experience them every day—one group of neurons is active while you are seeing colors, whereas the cells firing in another cortical neighborhood are associated with being in a jocular mood. When these neurons are stimulated by a neurosurgeon's electrode, the subject sees colors or erupts in laughter. Conversely, shutting down the brain during anesthesia eliminates these experiences.

IN BRIEF

Machines with human-level intelligence are on the horizon. **Whether** they will actually be conscious remains unknown. **Why?** Even the most sophisticated brain simulations are unlikely to produce conscious feelings.

Given these widely shared background assumptions, what will the evolution of true artificial intelligence imply about the possibility of artificial consciousness?

Contemplating this question, we inevitably come to a fork up ahead, leading to two fundamentally different destinations. The zeitgeist, as embodied in novels and movies such as *Blade Runner, Her* and *Ex Machina,* marches resolutely down the road toward the assumption that truly intelligent machines will be sentient; they will speak, reason, self-monitor and introspect. They are *eo ipso* conscious.

This path is epitomized most explicitly by the global neuronal workspace (GNW) theory, one of the dominant scientific theories of consciousness. The theory starts with the brain and infers that some of its peculiar architectural features are what gives rise to consciousness.

Its lineage can be traced back to the "blackboard architecture" of 1970s computer science, in which specialized programs accessed a shared repository of information, called the blackboard or central workspace. Psychologists postulated that such a processing resource exists in the brain and is central to human cognition. Its capacity is small, so only a single percept, thought or memory occupies the workspace at any one time. New information competes with the old and displaces it.

Cognitive neuroscientist Stanislas Dehaene and molecular biologist Jean-Pierre Changeux, both at the Collège de France in Paris, mapped these ideas onto the architecture of the brain's cortex, the outermost layer of gray matter. Two highly folded cortical sheets, one on the left and one on the right, each the size and thickness of a 14-inch pizza, are crammed into the protective skull. Dehaene and Changeux postulated that the workspace is instantiated by a network of pyramidal (excitatory) neurons linked to far-flung cortical regions, in particular the prefrontal, parietotemporal and midline (cingulate) associative areas.

Much brain activity remains localized and therefore unconscious—for example, that of the module that controls where the eyes look, something of which we are almost completely oblivious, or that of the module that adjusts the posture of our bodies. But when activity in one or more regions exceeds a threshold—say, when someone is presented with an image of a Nutella jar—it triggers an ignition, a wave of neural excitation that spreads throughout the neuronal workspace, brain-wide. That signaling therefore becomes available to a host of subsidiary processes such as language, planning, reward circuits, access to long-term memory, and storage in a short-term memory buffer. The act of globally broadcasting this information is what renders it conscious. The inimitable experience of Nutella is constituted by pyramidal neurons contacting the brain's motor-planning region—issuing an instruction to grab a spoon to scoop out some of the hazelnut spread. Meanwhile other modules transmit the message to expect a reward in the form of a dopamine rush caused by Nutella's high fat and sugar content.

Conscious states arise from the way the workspace algorithm processes the relevant sensory inputs, motor outputs, and internal variables related to memory, motivation and expectation. Global processing is what consciousness is about. GNW theory fully embraces the contemporary mythos of the near-infinite powers of computation. Consciousness is just a clever hack away.

INTRINSIC CAUSAL POWER

THE ALTERNATIVE PATH—integrated information theory (IIT)—takes a more fundamental approach to explaining consciousness.

Giulio Tononi, a psychiatrist and neuroscientist at the University of Wisconsin–Madison, is the chief architect of IIT, with others, myself included, contributing. The theory starts with experience and proceeds from there to the activation of synaptic circuits that determine the "feeling" of this experience. Integrated information is a mathematical measure quantifying how much "intrinsic causal power" some mechanism possesses. Neurons firing action potentials that affect the downstream cells they are wired to (via synapses) are one type of mechanism, as are electronic circuits, made of transistors, capacitances, resistances and wires.

Intrinsic causal power is not some airy-fairy ethereal notion but can be precisely evaluated for any system. The more its current state specifies its cause (its input) and its effect (its output), the more causal power it possesses.

IIT stipulates that any mechanism with intrinsic power, whose state is laden with its past and pregnant with its future, is conscious. The greater the system's integrated information, represented by the Greek letter Φ (a zero or positive number pronounced "fi"), the more conscious the system is. If something has no intrinsic causal power, its Φ is zero; it does not feel anything.

Given the heterogeneity of cortical neurons and their densely overlapping set of input and output connections, the amount of integrated information within the cortex is vast. The theory has inspired the construction of a consciousness meter currently under clinical evaluation, an instrument that determines whether people in persistent vegetative states or those who are minimally conscious, anesthetized or locked-in are conscious but unable to communicate or whether "no one is home." In analyses of the causal power of programmable digital computers at the level of their metal components—the transistors, wires and diodes that serve as the physical substrate of any computation—the theory indicates that their intrinsic causal power and their Φ are minute. Furthermore, Φ is independent of the software running on the processor, whether it calculates taxes or simulates the brain.

Indeed, the theory proves that two networks that perform the same input-output operation but have differently configured circuits can possess different amounts of Φ. One circuit may have no Φ, whereas the other may exhibit high levels. Although they are identical from the outside, one network experiences something while its zombie impostor counterpart feels nothing. The difference is under the hood, in the network's internal wiring. Put succinctly, consciousness is about *being,* not about *doing*.

The difference between these theories is that GNW emphasizes the function of the human brain in explaining consciousness, whereas IIT asserts that it is the intrinsic causal powers of the brain that really matter.

The distinctions reveal themselves when we inspect the brain's connectome, the complete specification of the exact synaptic wiring of the entire nervous system. Anatomists have already mapped the connectomes of a few worms. They are working on the connectome for the fruit fly and are planning to tackle the mouse within the next decade. Let us assume that in the future it will be possible to scan an entire human brain, with its roughly 100 billion neurons and quadrillion synapses, at the ultrastructural level after its owner has died and then simulate the organ on some advanced computer, maybe a quantum machine. If the model is faithful enough, this simulation will wake up and behave like a digital simulacrum of the deceased person—speaking and accessing his or her memories, cravings, fears and other traits.

If mimicking the functionality of the brain is all that is needed to create consciousness, as postulated by GNW theory, the simulated person will be conscious, reincarnated inside a computer. Indeed, uploading the connectome to the cloud so people can live on in the digital afterlife is a common science-fiction trope.

IIT posits a radically different interpretation of this situation: the simulacrum will feel as much as the software running on a fancy Japanese toilet—nothing. It will act like a person but without any innate feelings, a zombie (but without any desire to eat human flesh)—the ultimate deepfake.

To create consciousness, the intrinsic causal powers of the brain are needed. And those powers cannot be simulated but must be part and parcel of the physics of the underlying mechanism.

To understand why simulation is not good enough, ask yourself why it never gets wet inside a weather simulation of a rainstorm or why astrophysicists can simulate the vast gravitational power of a black hole without having to worry that they will be swallowed up by spacetime bending around their computer. The answer: because a simulation does not have the causal power to cause atmospheric vapor to condense into water or to cause spacetime to curve! In principle, however, it would be possible to achieve human-level consciousness by going beyond a simulation to build so-called neuromorphic hardware, based on an architecture built in the image of the nervous system.

There are other differences besides the debates about simulations. IIT and GNW predict that distinct regions of the cortex constitute the physical substrate of specific conscious experiences, with an epicenter in either the back or the front of the cortex. This prediction and others are now being tested in a large-scale collaboration involving six labs in the U.S., Europe and China that has just received $5 million in funding from the Templeton World Charity Foundation.

Whether machines can become sentient matters for ethical reasons. If computers experience life through their own senses, they cease to be purely a means to an end determined by their usefulness to us humans. They become an end unto themselves.

Per GNW, they turn from mere objects into subjects—each exists as an "I"—with a point of view. This dilemma comes up in the most compelling *Black Mirror* and *Westworld* television episodes. Once computers' cognitive abilities rival those of humanity, their impulse to push for legal and political rights will become irresistible—the right not to be deleted, not to have their memories wiped clean, not to suffer pain and degradation. The alternative, embodied by IIT, is that computers will remain only supersophisticated machinery, ghostlike empty shells, devoid of what we value most: the feeling of life itself. SA

MORE TO EXPLORE

What Is Consciousness, and Could Machines Have It? Stanislas Dehaene, Hakwan Lau and Sid Kouider in *Science,* Vol. 358, pages 486–492; October 27, 2017.

The Feeling of Life Itself: Why Consciousness Is Widespread but Can't Be Computed. Christof Koch. MIT Press, 2019.

FROM OUR ARCHIVES

Is the Brain's Mind a Computer Program? John R. Searle; January 1990.

scientificamerican.com/magazine/sa