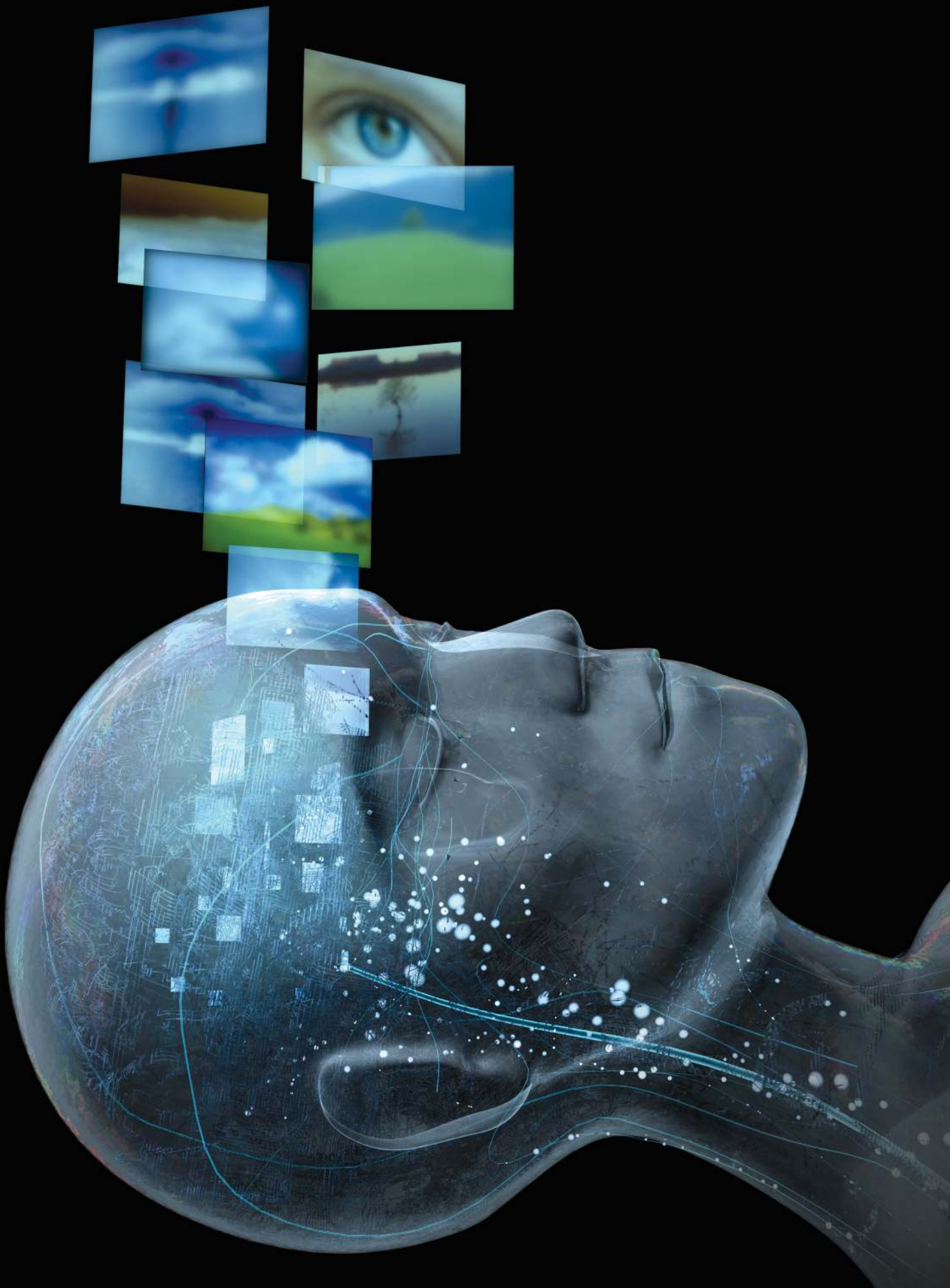


Can Machines Be Conscious?

YES—AND A NEW TURING TEST MIGHT PROVE IT

BY CHRISTOF KOCH AND GIULIO TONONI

WOULD YOU SELL your soul on eBay? Right now, of course, you can't. But in some quarters it is taken for granted that within a generation, human beings—including you, if you can hang on for another 30 years or so—will have an alternative to death: being a ghost in a machine. You'll be able to upload your mind—your thoughts, memories, and personality—to a computer. And once you've reduced your consciousness to patterns of electrons, others will be able to copy it, edit it, sell it, or pirate it. It might be bundled with other electronic minds. And, of course, it could be deleted.

That's quite a scenario, considering that at the moment, nobody really knows exactly what consciousness is. Pressed for a pithy definition, we might call it the ineffable and enigmatic inner life of the mind. But that hardly captures the whirl of thought and sensation that blossoms when you see a loved one after a long absence, hear an exquisite violin solo, or relish an incredible meal. Some of the most brilliant minds in human history have pondered consciousness, and after a few thousand years we still can't say for sure if it is an intangible phenomenon or maybe even a kind of substance different from matter. We know it arises in the brain, but we don't know how or where in the brain. We don't even know if it requires specialized brain cells (or neurons) or some sort of special circuit arrangement of them.

Nevertheless, some in the singularity crowd are confident that we are within a few decades of building a computer, a simulacrum, that can experience the color red, savor the smell of a rose, feel pain and pleasure, and fall in love. It might be a robot with a "body." Or it might just be software—a huge, ever-changing cloud of bits that inhabit an immensely complicated and elaborately constructed virtual domain.

We are among the few neuroscientists who have devoted a substantial part of their careers to studying consciousness. Our work has given us a unique perspective on what is arguably the most momentous issue in all of technology: whether consciousness will ever be artificially created.

We think it will—eventually. But perhaps not in the way that the most popular scenarios have envisioned it.

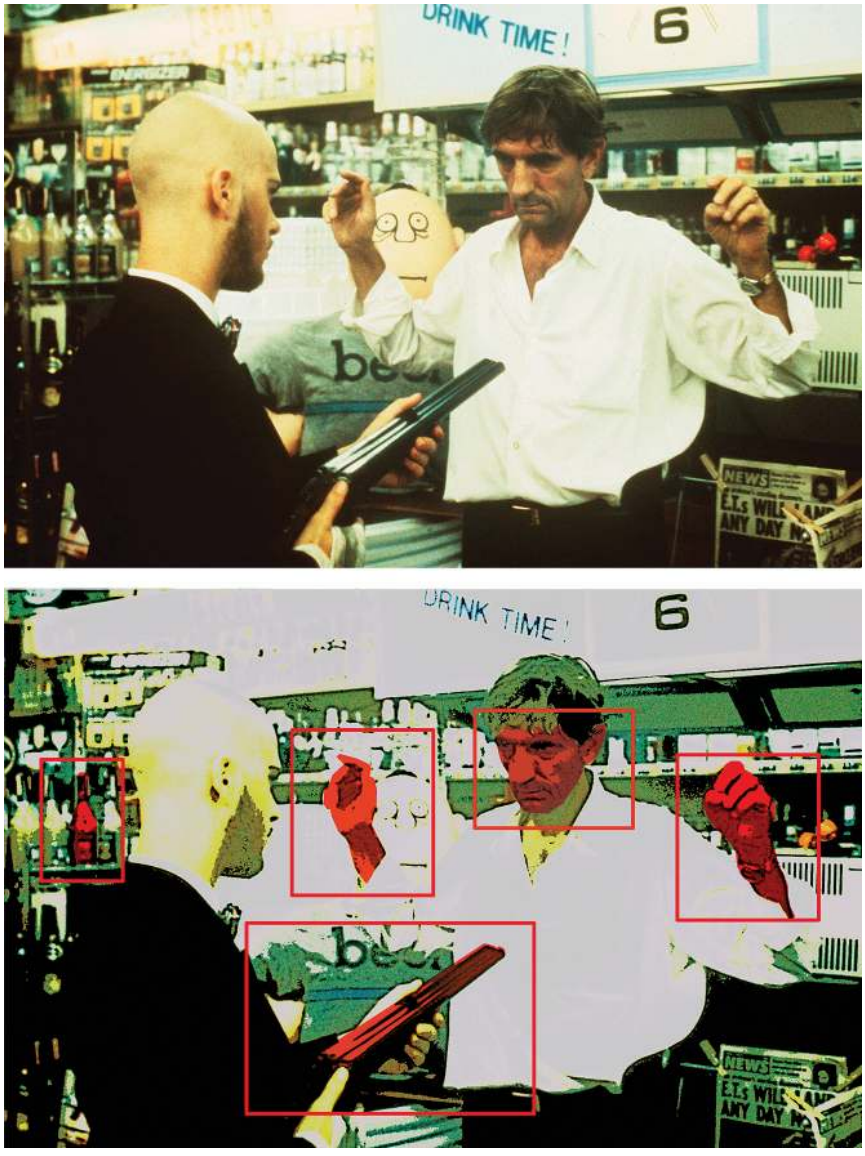
CONSCIOUSNESS IS PART of the natural world. It depends, we believe, only on mathe-

BRYAN CHRISTIE DESIGN

matics and logic and on the imperfectly known laws of physics, chemistry, and biology; it does not arise from some magical or otherworldly quality. That's good news, because it means there's no reason why consciousness can't be reproduced in a machine—in theory, anyway.

In humans and animals, we know that the specific content of any conscious experience—the deep blue of an alpine sky, say, or the fragrance of jasmine redolent in the night air—is furnished by parts of the cerebral cortex, the outer layer of gray matter associated with thought, action, and other higher brain functions. If a sector of the cortex is destroyed by stroke or some other calamity, the person will no longer be conscious of whatever aspect of the world that part of the brain represents. For instance, a person whose visual cortex is partially damaged may be unable to recognize faces, even though he can still see eyes, mouths, ears, and other discrete facial features. Consciousness can be lost entirely if injuries permanently damage most of the cerebral cortex, as seen in patients like Terri Schiavo, who suffered from persistent vegetative state. Lesions of the cortical white matter, containing the fibers through which parts of the brain communicate, also cause unconsciousness. And small lesions deep within the brain along the midline of the thalamus and the midbrain can inactivate the cerebral cortex and indirectly lead to a coma—and a lack of consciousness.

To be conscious also requires the cortex and thalamus—the corticothalamic system—to be constantly suffused in a bath of substances known as neuromodulators, which aid or inhibit the transmission of nerve impulses. Finally, whatever the mechanisms necessary for consciousness, we know they must exist in both cortical hemispheres independently.



A BETTER TURING TEST: Shown this frame from the cult classic *Repo Man* [top], a conscious machine should be able to home in on the key elements [bottom]—a man with a gun, another man with raised arms, bottles on shelves—and conclude that it depicts a liquor-store robbery.
PHOTO: EDGE CITY/UNIVERSAL/THE KOBAL COLLECTION

Much of what goes on in the brain has nothing to do with being conscious, however. Widespread damage to the cerebellum, the small structure at the base of the brain, has no effect on consciousness, despite the fact that more neurons reside there than in any other part of the brain. Neural activity obviously plays some essential role in consciousness but in itself is not enough to sustain a conscious state. We know that at the beginning of a deep sleep, consciousness fades, even though the neurons in the corticothalamic system continue to fire at a level of activity similar to that of quiet wakefulness.

Data from clinical studies and from basic research laboratories, made possible by the use of sophisticated instruments that detect and record neuronal activity, have given us a complex if still rudimentary understanding of the myriad processes that give rise to consciousness. We are still a *very* long way from being able to use this knowledge to build a conscious machine. Yet we can already take the first step in that long journey: we can list some aspects of consciousness that are not strictly necessary for building such an artifact.

Remarkably, consciousness does not seem to require many of the things we associate most deeply with being human: emotions, memory, self-reflection, language, sensing the world, and acting in it. Let's start with sensory input and motor output: *being conscious requires neither*. We humans are generally aware of what goes on around us and occasionally of what goes on within our own bodies. It's only natural to infer that consciousness is linked to our interaction with the world and with ourselves.

Yet when we dream, for instance, we are virtually disconnected from the environment—we acknowledge almost nothing of what happens around us, and our muscles are largely paralyzed. Nevertheless, we are conscious, sometimes vividly and grippingly so. This mental activity is reflected in electrical recordings of the dreaming brain showing that the cortico-

thalamic system, intimately involved with sensory perception, continues to function more or less as it does in wakefulness.

Neurological evidence points to the same conclusion. People who have lost their eyesight can both imagine and dream in images, provided they had sight earlier in their lives. Patients with locked-in syndrome, which renders them almost completely paralyzed, are just as conscious as healthy subjects. Following a debilitating stroke, the French editor Jean-Dominique Bauby dictated his memoir, *The Diving Bell and the Butterfly*, by blinking his left eye. Stephen Hawking is a world-renowned physicist, best-selling author, and occasional guest star on "The Simpsons," despite being immobilized from a degenerative neurological disorder.

So although being conscious depends on brain activity, it does not require any interaction with the environment. Whether the *development* of consciousness requires such interactions in early childhood, though, is a different matter.

How about emotions? Does a conscious being need to feel and display them? No: *being conscious does not require emotion*. People who've suffered damage to the frontal area of the brain, for instance, may exhibit a flat, emotionless affect; they are as dispassionate about their own predicament as they are about the problems of people around them. But even though their behavior is impaired and their judgment may be unsound, they still experience the sights and sounds of the world much the way normal people do.

Primal emotions like anger, fear, surprise, and joy are useful and perhaps even essential for the survival of a conscious organism. Likewise, a conscious machine might rely on emotions to make choices and deal with the complexities of the world. But it could be just a cold, calculating engine—and yet still be conscious.

Psychologists argue that consciousness requires selective attention—that is, the ability to focus on a given object, thought, or activity. Some have even argued that consciousness *is* selective attention. After all, when you pay attention to something, you become conscious of that thing and its properties; when your attention shifts, the object fades from consciousness.

Nevertheless, recent evidence favors the idea that a person can consciously perceive an event or object without paying attention to it. When you're focused on a riveting movie, your surroundings aren't reduced to a tunnel. You may not hear the phone ringing or your spouse calling your name, but you remain aware of certain aspects of the world around you. And here's a surprise: the converse is also true. People can attend to events or objects—that is, their brains can preferentially process them—without consciously perceiving them. This fact suggests that *being conscious does not require attention*.

One experiment that supported this conclusion found that, as strange as it sounds, people could pay attention to an object that they never "saw." Test subjects were shown static images of male and female nudes in one eye and rapidly flashing colored squares in the other eye. The flashing color rendered the nudes invisible—the subjects couldn't even say where the nudes were in the image. Yet the psychologists showed that subjects nevertheless registered the unseen image if it was of the opposite sex.

What of memory? Most of us vividly remember our first kiss, our first car, or the images of the crumbling Twin Towers on 9/11. This kind of episodic memory would seem to be an integral part of consciousness. But the clinic tells us otherwise: *being conscious does not require either explicit or working memory*.

In 1953, an epileptic man known to the public only as H.M. had most of his hippocampus and neighboring regions on both sides of the brain surgically removed as an experimental treatment for his condition. From that day on, he couldn't acquire any new long-term memories—not of the nurses and doctors who treated him, his room at the hospital, or any unfamiliar well-wishers who dropped by. He could recall only events that happened before his surgery. Such impairments, though, didn't turn H.M. into a zombie. He is still alive today, and even if he can't remember events from one day to the next, he is without doubt conscious.

The same holds true for the sort of working memory you need to perform any number of daily activities—to dial a phone number you just looked up or measure out the correct amount of crushed thyme given in the cookbook you just consulted. This memory is called dynamic because it lasts only as long as neuronal circuits remain active. But as with long-term memory, you don't need it to be conscious.

Self-reflection is another human trait that seems deeply linked to consciousness. To assess consciousness, psychologists and other scientists often rely on verbal reports from their subjects. They ask questions like "What did you see?" To answer, a subject conjures up an image by "looking inside" and recalling whatever it was that was just viewed. So it is only natural to suggest that consciousness arises through your ability to reflect on your perception.

As it turns out, though, *being conscious does not require self-reflection*. When we become absorbed in some intense perceptual task—such as playing a fast-paced video game, swerving on a motorcycle through moving traffic, or running along a mountain trail—we are vividly conscious of the external world, without any need for reflection or introspection.

Neuroimaging studies suggest that we can be vividly conscious even when the front of the cerebral cortex, involved in judgment and self-representation, is relatively inactive. Patients with widespread injury to the front of the brain demonstrate serious deficits in their cognitive, executive, emotional, and planning abilities. But they appear to have nearly intact perceptual abilities.

Finally, *being conscious does not require language*. We humans affirm our consciousness through speech, describing and discussing our experiences with one another. So it's natural to think that speech and consciousness are inextricably linked. They're not. There are many patients who lose the ability to understand or use words and yet remain conscious. And infants, monkeys, dogs, and mice cannot speak, but they are conscious and can report their experiences in other ways.

So what about a machine? We're going to assume that a machine does not require anything to be conscious that a naturally evolved organism—you or me, for example—doesn't require. If that's the case, then, to be conscious a machine does not need to engage with its environment, nor does it need long-term memory or working memory; it does not require attention, self-reflection, language, or emotion. Those things may help the machine survive in the real world. But to simply have subjective experience—being pleased at the sight of wispy white clouds scurrying across a perfectly blue sky—those traits are probably not necessary.

So what *is* necessary? What are the essential properties of consciousness, those without which there is no experience whatsoever?

We think the answer to that question has to do with the amount of *integrated information* that an organism, or a machine, can generate. Let's say you are

facing a blank screen that is alternately on or off, and you have been instructed to say "light" when the screen turns on and "dark" when it turns off. Next to you, a photodiode—one of the very simplest of machines—is set up to beep when the screen emits light and to stay silent when the screen is dark. The first problem that consciousness poses boils down to this: both you and the photodiode can differentiate between the screen being on or off, but while you can see light or dark, the photodiode does not consciously "see" anything. It merely responds to photons.

The key difference between you and the photodiode has to do with how much information is generated when the differentiation between light and dark is made. Information is classically defined as the reduction of uncertainty that occurs when one among many possible outcomes is chosen. So when the screen turns dark, the photodiode enters one of its two possible states; here, a state corresponds to one bit of information. But when *you* see the screen turn dark, you enter one out of a huge number of states: seeing a dark screen means you aren't seeing a blue, red, or green screen, the Statue of Liberty, a picture of your child's piano recital, or any of the other uncountable things that you have ever seen or could ever see. To you, "dark" means not just the opposite of light but also, and simultaneously, something different from colors, shapes, sounds, smells, or any mixture of the above.

So when you look at the dark screen, you rule out not just "light" but countless other possibilities. You don't think of the stupefying number of possibilities, of course, but their mere existence corresponds to a huge amount of information.

Conscious experience consists of more than just differentiating among many states, however. Consider an idealized 1-megapixel digital camera. Even if each photodiode in the imager were just binary, the number of different patterns that imager could record is $2^{1\,000\,000}$. Indeed, the camera could easily enter a different state for every frame from every movie that was or could ever be produced. It's a staggering amount of information. Yet the camera is obviously not conscious. Why not?

We think that the difference between you and the camera has to do with *integrated* information. The camera can indeed be in any one of an absurdly large number of different states. However, the 1-megapixel sensor chip isn't a single integrated system but rather a collection of one million individual, completely independent photodiodes, each with a repertoire of two states. And a million photodiodes are collectively no smarter than one photodiode.

By contrast, the repertoire of states available to you cannot be subdivided. You know this from experience: when you consciously see a certain image, you experience that image as an integrated whole. No matter how hard you try, you cannot divvy it up into smaller thumbprint images, and you cannot experience its colors independently of the shapes, or the left half of your field of view independently of the right half. Underlying this unity is a multitude of causal interactions among the relevant parts of your brain. And unlike chopping up the photodiodes in a camera sensor, disconnecting the elements of your brain that feed into consciousness would have profoundly detrimental effects.

TO BE CONSCIOUS, then, you need to be a single integrated entity with a large repertoire of states. Let's take this one step further: your *level* of consciousness has to do with how much integrated information you can generate. That's why you have a higher level of consciousness than a tree frog or a supercomputer.

It is possible to work out a theoretical framework for gauging how effective different neural architectures would be at generating integrated information and therefore attaining a conscious state. This framework, the integrated information theory of consciousness, or IIT, is grounded in the mathematics of information and complexity theory and provides a specific measure of the amount of integrated information generated by any system comprising interacting parts. We call that measure Φ and express it in bits. The larger the value of Φ, the larger the entity's conscious repertoire. (For students of information theory, Φ is an intrinsic property of the system, and so it is different from the Shannon information that can be sent through a channel.)

IIT suggests a way of assessing consciousness in a machine—a Turing Test for consciousness, if you will. Other attempts at gauging machine consciousness, or at least intelligence, have fallen short. Carrying on an engaging conversation in natural language or playing strategy games were at various times thought to be uniquely human attributes. Any machine that had those capabilities would also have a human intellect, researchers once thought. But subsequent events proved them wrong—computer programs such as the chatterbot ALICE and the chess-playing supercomputer Deep Blue, which famously bested Garry Kasparov in 1997, demonstrated that machines can display human-level performance in narrow tasks. Yet none of those inventions displayed evidence of consciousness.

Scientists have also proposed that displaying emotion, self-recognition, or purposeful behavior are suitable criteria for machine consciousness. However, as we mentioned earlier, there are people who are clearly conscious but do not exhibit those traits.

What, then, would be a better test for machine consciousness? According to IIT, consciousness implies the availability of a large repertoire of states belonging to a single integrated system. To be useful, those internal states should also be highly informative about the world.

One test would be to ask the machine to describe a scene in a way that efficiently differentiates the scene's key features from the immense range of other possible scenes. Humans are fantastically good at this: presented with a photo, a painting, or a frame from a movie, a normal adult can describe what's going on, no matter how bizarre or novel the image is.

Consider the following response to a particular image: "It's a robbery—there's a man holding a gun and pointing it at another man, maybe a store clerk." Asked to elaborate, the person could go on to say that it's probably in a liquor store, given the bottles on the shelves, and that it may be in the United States, given the English-language newspaper and signs. Note that the exercise here is not to spot as many details as one can but to discriminate the scene, as a whole, from countless others.

So this is how we can test for machine consciousness: show it a picture and ask it for a concise description [see photos, "A Better Turing Test"]. The machine should

> Consciousness does not seem to require many of the things we associate with being human

be able to extract the gist of the image (it's a liquor store) and what's happening (it's a robbery). The machine should also be able to describe which objects are in the picture and which are not (where's the getaway car?), as well as the spatial relationships among the objects (the robber is holding a gun) and the causal relationships (the other man is holding up his hands because the bad guy is pointing a gun at him).

The machine would have to do as well as any of us to be considered as conscious as we humans are—so that a human judge could not tell the difference—and not only for the robbery scene but for any and all other scenes presented to it.

No machine or program comes close to pulling off such a feat today. In fact, image understanding remains one of the great unsolved problems of artificial intelligence. Machine-vision algorithms do a reasonable job of recognizing ZIP codes on envelopes or signatures on checks and at picking out pedestrians in street scenes. But deviate slightly from these well-constrained tasks and the algorithms fail utterly.

Very soon, computer scientists will no doubt create a program that can automatically label thousands of common objects in an image—a person, a building, a gun. But that software will still be far from conscious. Unless the program is explicitly written to conclude that the combination of man, gun, building, and terrified customer implies "robbery," the program won't realize that something dangerous is going on. And even if it were so written, it might sound a false alarm if a 5-year-old boy walked into view holding a toy pistol. A sufficiently conscious machine would not make such a mistake.

WHAT IS THE best way to build a conscious machine? Two complementary strategies come to mind: either copying the mammalian brain or evolving a machine. Research groups worldwide are already pursuing both strategies, though not necessarily with the explicit goal of creating machine consciousness.

Though both of us work with detailed biophysical computer simulations of the cortex, we are not optimistic that modeling the brain will provide the insights needed to construct a conscious machine in the next few decades. Consider this sobering lesson: the roundworm *Caenorhabditis elegans* is a tiny creature whose brain has 302 nerve cells. Back in 1986, scientists used electron microscopy to painstakingly map its roughly 6000 chemical synapses and its complete wiring diagram. Yet more than two decades later, there is still no working model of how this minimal nervous system functions.

Now scale that up to a human brain with its 100 billion or so neurons and a couple hundred trillion synapses. Tracing all those synapses one by one is close to impossible, and it is not even clear whether it would be particularly useful, because the brain is astoundingly plastic, and the connection strengths of synapses are in constant flux. Simulating such a gigantic neural network model in the hope of seeing consciousness emerge, with millions of parameters whose values are only vaguely known, will not happen in the foreseeable future.

A more plausible alternative is to start with a suitably abstracted mammal-like architecture and evolve it into a conscious entity. Sony's robotic dog, Aibo, and its humanoid, Qrio, were rudimentary attempts; they operated under a large number of fixed but flexible rules. Those rules yielded some impressive, lifelike behavior—chasing balls, dancing, climbing stairs—but such robots have no chance of passing our consciousness test.

So let's try another tack. At MIT, computational neuroscientist Tomaso Poggio has shown that vision systems based on hierarchical, multilayered maps of neuronlike elements perform admirably at learning to categorize real-world images. In fact, they rival the performance of state-of-the-art machine-vision systems. Yet such systems are still very brittle. Move the test setup from cloudy New England to the brighter skies of Southern California and the system's performance suffers. To begin to approach human behavior, such systems must become vastly more robust; likewise, the range of what they can recognize must increase considerably to encompass essentially all possible scenes.

Contemplating how to build such a machine will inevitably shed light on scientists' understanding of our own consciousness. And just as we ourselves have evolved to experience and appreciate the infinite richness of the world, so too will we evolve constructs that share with us and other sentient animals the most ineffable, the most subjective of all features of life: consciousness itself. ❑

TO PROBE FURTHER For more on the integrated information theory of consciousness, go to http://spectrum.ieee.org/jun08/consciousmachines.

CHRIS MEYER/INDIANA UNIVERSITY

EXPERT VIEW:

Douglas Hofstadter

WHO HE IS
Pioneer in computer modeling of mental processes; director of the Center for Research on Concepts and Cognition at Indiana University, Bloomington; winner of the 1980 Pulitzer Prize for general nonfiction.

SINGULARITY WILL OCCUR
Someday in the distant future

MACHINE CONSCIOUSNESS WILL OCCUR
Yes

MOORE'S LAW WILL CONTINUE FOR
20 more years

THOUGHTS
"It might happen someday, but I think life and intelligence are far more complex than the current singularitarians seem to believe, so I doubt it will happen in the next couple of centuries. [The ramifications] will be enormous, since the highest form of sentient beings on the planet will no longer be human. Perhaps these machines—our 'children'—will be vaguely like us and will have culture similar to ours, but most likely not. In that case, we humans may well go the way of the dinosaurs."