NEUROSCIENCE

A Test for Consciousness

How will we know when we've built a sentient computer? By making it solve a simple puzzle

By Christof Koch and Giulio Tononi

COMPUTERS INCH EVER CLOSER TO BEHAVING LIKE INTELLIGENT HUMAN BEINGS—WITNESS the ability of IBM's Watson to beat the all-time champs of the television quiz show *Jeopardy*. So far, though, most people would doubt that computers truly "see" a visual scene full of shapes and colors in front of their cameras, that they truly "hear" a question through their microphones, that they feel anything—experience consciousness—the way humans do, despite computers' remarkable ability to crunch data at superhuman speed.

How would we know if a machine had taken on this seemingly ineffable quality of conscious awareness? Our strategy relies on the knowledge that only a conscious machine can demonstrate a subjective understanding of whether a scene depicted in some ordinary photograph is "right" or "wrong." This ability to assemble a set of facts into a picture of reality that makes eminent sense—or know, say, that an elephant should not be perched on top of the Eiffel Tower—defines an essential property of the conscious mind. A roomful of IBM supercomputers, in contrast, still cannot fathom what makes sense in a scene.

Understanding the attributes of a sentient machine will allow humans not only to understand the workings of our own brains but to prepare for that day, envisaged in science fiction, when we must learn to live with another form of conscious being that we ourselves created. This understanding may even allow us to address one of the most profound questions that has beset philosophers throughout the ages: What is consciousness?

IS IT MAN OR GOLEM?

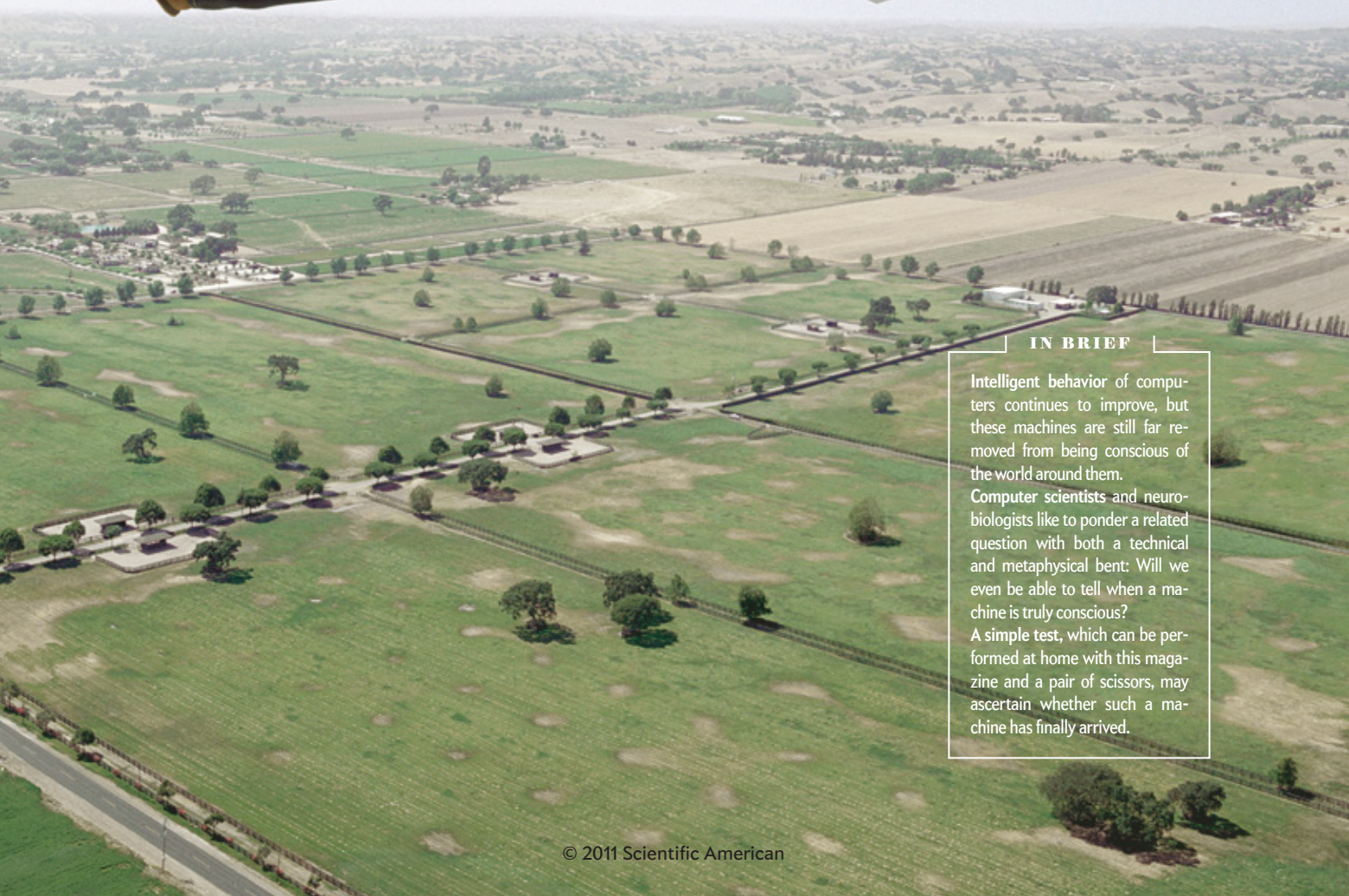
PHILOSOPHERS HAVE LONG PONDERED the question of whether a man-made simulacrum, be it the mythical golem or a machine in a box, can feel or experience anything. Then, in 1950, Alan Turing, the British mathematician who helped to break the Enigma code used by the feared Nazi submarine force in World War II, published a paper that launched the field of artificial intelligence. In an article in the journal *Mind,* Turing proposed replacing the impossi-

GEOF KERN

What's wrong with this picture? To judge that this image is incorrect, a machine would need to be conscious of many things about the world (unless programmed for just such a photograph).

Christof Koch is Lois and Victor Troendle Professor of Cognitive and Behavioral Biology at the California Institute of Technology and chief scientific officer at the Allen Institute for Brain Science in Seattle. He serves on *Scientific American Mind*'s board of advisers.

Giulio Tononi is David P. White Chair in Sleep Medicine and a Distinguished Professor in Consciousness Science at the University of Wisconsin–Madison.



IN BRIEF

Intelligent behavior of computers continues to improve, but these machines are still far removed from being conscious of the world around them.

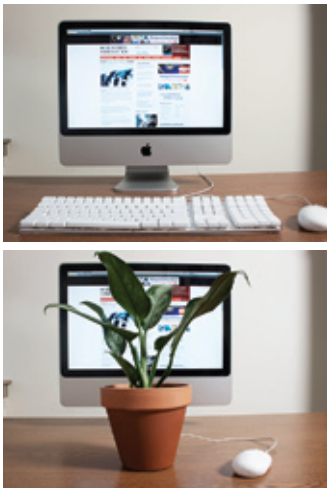
Computer scientists and neurobiologists like to ponder a related question with both a technical and metaphysical bent: Will we even be able to tell when a machine is truly conscious?

A simple test, which can be performed at home with this magazine and a pair of scissors, may ascertain whether such a machine has finally arrived.

bly vague question—Can machines think?—with a much more practical one—Can we build machines that, when queried via Teletype, cannot be distinguished from a human?

The version of the Turing test employed today has a human judge interacting via a computer screen with a human or a software program in the "natural language" that we use to communicate. The conversation between the judge and his or her partner can address any topic. If after some suitable interval, the judge cannot be sure whether the partner is human, at the very least it can be said to be as intelligent as a person, having passed the Turing test. Over the years chatterbots—conversational programs designed to simulate intelligent small talk—have, on occasion, deceived judges, but not for long.

The two of us come to the question of machine consciousness not as computer scientists but as neurobiologists interested in how brains give rise to subjective experience. We probe the brains of volunteers or patients with neurological disorders in magnetic scanners or record their brain waves with electroencephalography. We also carry out similar investigations of the brains of rodents and other animals. In doing so, we and many of our colleagues are homing in on the so-called neuronal correlates of consciousness: the minimal brain mechanisms that together suffice to cause any specific conscious sensation, such as observing a garish, orange sunset. Yet what the field has lacked until recently is a general theory that allows us to assess, in a principled way, whether a brain-injured patient, a fetus, a mouse or a silicon simulacrum can experience conscious sensations.



This not that: A test for consciousness could ask a nominally sentient machine which of two pictures are wrong, a task that would stump any present-day automaton.

What we call the integrated information theory of consciousness provides one way to tackle that challenge. It touches on a critical determinant of consciousness. Many people have an intuitive understanding that the subjective, phenomenal states that make up everyday experience—the way each of us experiences a smell, a visual scene, a thought or a recollection in a highly individual manner—must somehow relate to how the brain integrates incoming sensory signals with information from memory into a cohesive picture of the world. But how can this intuition be made more precise?

The integrated information theory addresses this need by putting forth two axioms. First, consciousness is highly informative. This is because each particular conscious state, when it occurs, rules out an immense number of other possible states, from which it differs in its own particular way. Think of all the frames from all the movies you have ever seen. Each frame, each view, is a specific conscious percept: when you perceive that frame, your brain rules out trillions of other possible images. Even after awakening in a dark room, seemingly the simplest visual experience, the percept of pitch-blackness implies that you do not see a well-lit living room, the intricate canopy of the jungle or any of countless other scenes that could present themselves to the mind.

Second, conscious information is integrated. When you become conscious of your friend's face, you cannot fail to notice that she is crying and wearing glasses. No matter how hard you try, you cannot separate the left half of your field of view from the right or switch to seeing things in black and white. Whatever scene enters consciousness remains whole and complete; it cannot be subdivided into independent and unrelated components that can be experienced on their own.

The unified nature of consciousness stems from a multitude of interactions among relevant parts of your brain. If areas of the brain become disconnected, as occurs in anesthesia or in deep sleep—consciousness wanes and perhaps disappears.

To be conscious, then, you need to be a single, integrated entity with a large repertoire of distinguishable states—the definition of information. A system's capacity for integrated information, and thus for consciousness, can be measured by asking how much information a system contains above and beyond that possessed by its individual parts. This quantity, called Φ, or phi (pronounced "fi"), can be calculated, in principle, for any system, whether it be a brain, a robot or a manually adjustable thermostat. Think of Φ as the irreducibility of a system to a mere collection of parts, measured in bits. For the level of Φ and consciousness to be high, a system must be made of parts that are specialized and well integrated—parts that do more together than they can alone.

If the elements of a system are largely independent, like the sensors in a digital camera or the bits in a computer's memory, Φ will be low. It will also be low if the elements all do the same thing because they are not specialized and are therefore redundant; Φ also stays low if the elements of a system interconnect at random. But for certain parts of the brain, such as the cerebral cortex—where neurons are richly endowed with specific connections—Φ will be high. This measure of a system's integration can also apply to silicon circuits encased in a metal box. With sufficiently complex connections among the transistors and memory elements, computers, as with the brain, would reach high levels of integrated information.

Other than measuring Φ from the machine's wiring—a difficult task—how can we know whether a machine is sentient? What is a practical test? One way to probe for information integration would be to ask it to perform a task that any six-year-old can ace: "What's wrong with this picture?" Solving that simple problem requires having lots of contextual knowledge, vastly more than can be supplied with the algorithms that advanced computers depend on to identify a face or detect credit-card fraud.

Pictures of objects or natural scenes consist of massively intricate relations among pixels and objects—hence the adage "a picture is worth a thousand words." The evolution of our visual system, our neurological development during childhood and a lifetime of experience enable us to instantly know whether all the components fit together properly: Do the textures, depths, colors, spatial relations among the parts, and so on, make sense?

A computer that analyzes an image—to see that the information in it does not cohere—requires far more processing than do linguistic queries of a computer database. Computers may have beaten humans at sophisticated games, but they still

ERIC WEEKS

lack the ability to answer arbitrary questions about what is going on in a photograph. The degree of information integration explains why. Although the hard disk in a modern computer exceeds the capacity of our lifetime of memories, that information remains unintegrated: each element of the system stays largely disconnected from the others.

SEE-THROUGH COWS

TAKE JUST ONE EXAMPLE, a photograph of your desk in your iPhoto library. Your computer does not know whether, amid the usual clutter on your desk, your iMac on the left and your iPad on the right make sense together. Worse, the computer does not know that while the iMac and the iPad go together well, a potted plant instead of the keyboard is simply weird; or that it is impossible for the iPad to float above the table; or that the right side of the photograph fits well with the left side, whereas the right side of a multitude of other photographs would be wrong. To your computer, all pixels are just a vast, disconnected tapestry of three numbers (corresponding to three colors), with no particular meaning. To you, an image is meaningful because it is chock-full of connections among its parts, at many levels, ranging from pixels to objects to scenes. And these relations not only specify which parts of the image go well together but which ones do not. According to our theory, this integrated web of related knowledge gives each image an identity by distinguishing it from myriad others and imbues you with the capacity to become conscious of the world.

The same integration would also tell even a six-year-old that many incongruous pictures are ridiculous: an ice-skater on a rug in the living room, a transparent cow or a cat chasing a dog. And therein lies the secret of determining whether a computer is conscious. These obvious violations of our expectations testify to the remarkable knowledge we have of the way in which certain events and objects occur together, but the vast majority do not.

Testing a computer's understanding of an image does not require the conventional Turing test protocol of typing in a query to a machine. Instead you can simply pick some images at random from the Web. Black out a strip running vertically down the central third of each one, then shuffle the remaining left and right sides of the pictures. The parts of the composites will not match, except in one case, in which the left side is evidently from the same picture as the right side. The computer would be challenged to select the one picture that is correct. The black strip in the middle prevents the use of simple image-analysis strategies that computers use today—say, matching lines of texture or color across the separated, partial images. The split-image test requires a high level of visual understanding and the ability to deduce how the pieces of the image fit together.

Another test inserts objects into several images so that these objects make sense in each except for one, and the computer must detect the odd one out. A hammer on a workbench belongs there, but a tool is never suspended in midair. And a keyboard placed in front of an iMac is the right choice, not a potted plant.

A variety of computer strategies that rely on matching low-level statistical data of image characteristics such as color, edges or texture might manage to defeat one of these tests, but presenting many different image tests would defeat today's machines. The specifics of the tests that would actually be of practical use require more work. This exercise, though, highlights the enormous amount of integrated knowledge that you perceive consciously and throws into sharp relief the very narrow and highly specialized knowledge possessed by current machine-vision systems. Yes, today's machines can pick out the face of a likely terrorist from a database of a million faces, but they will not know his age, gender or ethnicity, whether he is looking directly at the viewer or not, or whether he is frowning or smiling. And they will not know that if he is shaking hands with George Washington, the photograph is probably digitally doctored. Any conscious human can apprehend all these things and more in a single glance.

Knowing all this, what can we expect for the near future? To the extent that a particular task can be singled out and characterized in isolation from other tasks, it can be taken over by machines. Fast algorithms can rapidly search through huge databases and beat humans at chess and *Jeopardy*. Sophisticated machine-learning algorithms can be trained to recognize faces or detect pedestrians faster and better than we do by exposing the computer to a large number of relevant examples labeled by humans. We can easily envision scenarios in which increasingly specialized tasks will be relegated to machines. Advanced computer-vision systems are coming of age, and in less than a decade a robust and largely autonomous driving mode will become an option.

And yet we predict that such machine-vision systems will not answer a simple question about the scene in front of the car: Does the Chicago skyline, seen at a distance from the approaching highway, resemble a burned tree grove emerging from the mist? And it will not realize that a giant banana next to the gas station would be out of place (except perhaps in Los Angeles). Answering such questions—and million of others—or spotting what is wrong with the banana would require countless dedicated software modules that no one could build in anticipation of that particular question. If we are right, although advanced machine-vision systems based on a set of specialized, parallel modules will make driving largely automatic—and will similarly simplify many other daily tasks—these systems will not consciously see a scene ahead.

Yet a different kind of machine can be envisioned, too—one in which knowledge of the innumerable relations among the things in our world is embodied in a single, highly integrated system. In such a machine, the answer to the question "What's wrong with this picture?" would pop out because whatever is awry would fail to match some of the intrinsic constraints imposed by the way information is integrated within a given system.

Such a machine would be good at dealing with things not easily separable into independent tasks. Based on its ability to integrate information, it would consciously perceive a scene. And we suspect that to achieve high levels of integration, such a machine might well exploit the structural principles in the mammalian brain. These machines will easily pass the tests we have described, and when they do they will share with us the gift of consciousness—this most enigmatic feature of the universe. SA

MORE TO EXPLORE

Can Machines Be Conscious? Christof Koch and Giulio Tononi in *IEEE Spectrum*, Vol. 45, No. 6, pages 54–59; June 2008.

Consciousness as Integrated Information: A Provisional Manifesto. Giulio Tononi in *Biological Bulletin*, Vol. 215, No. 3, pages 216–242; December 2008.

SCIENTIFIC AMERICAN ONLINE

Join the magazine's Fool-the-Machine contest at
ScientificAmerican.com/jun2011/koch-contest