scientificamerican.com

Testing for Consciousness in Machines

Giulio Tononi, Christof Koch

7-9 minutes

Credit: Dan Burn-Forti *Getty Images*

HOW WOULD WE KNOW if a machine is conscious? As computers inch closer to human-level performance—witness IBM's Watson victory over the all-time champs of the television quiz show *Jeopardy*—this question is becoming more pressing. So far, though, despite their ability to crunch data at superhuman speed, we suspect that unlike us, computers do not truly "see" a visual scene full of shapes and colors in front of their cameras; they don't "hear" a question through their microphones; they don't feel anything. Why do we think so, and how could we test if they do or do not experience a scene the way we do?

Consciousness, we have suggested, has two fundamental properties [see the July/August 2009 column by Christof Koch, "A Theory of Consciousness"]. First, every experience is highly informative. Any particular conscious state rules out an immense number of other possible states, from which it differs in its own particular way. Even the simple percept of pitch-blackness implies 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: think of all the frames from all the movies you have ever seen.

Second, conscious information is integrated. 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 your consciousness remains whole and complete: it cannot be subdivided into unrelated components that can be experienced on their own. Each experience, then, is a whole that acquires its meaning by how it can be distinguished from countless others, based on a lot of knowledge about the world. Our brain, with its multitude of specialized but interacting parts, seems optimally

adapted to achieving this feat of information integration. Indeed, if the relevant parts of our cerebral cortex become disconnected, as occurs in anesthesia or in deep sleep—consciousness wanes and perhaps disappears.

What's Wrong?

If consciousness requires this ability to generate an integrated picture that incorporates a lot of knowledge about the world, how could we know whether a computer is sentient? What is a practical test?

As we propose in the June 2011 issue of *Scientific American*, one way to probe for information integration would be to ask the computer to perform a task that any six-year-old child 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.

Views of objects or natural scenes consist of massively intricate relations among pixels and objects—hence the adage "a picture is worth a thousand words." Analyzing an image to see that something does not makes sense requires far more processing than do linguistic queries of a computer database. Computers may have beaten humans at sophisticated games, but they still lack an ability to answer arbitrary questions about what is going on in a photograph. In contrast, our visual system, thanks to its evolutionary history, its development during childhood and a lifetime of experience, enables 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?

Take just one example, a photograph of your workspace. Unless it is specifically programmed for that purpose, a computer analyzing the scene would not know whether, amid the usual clutter on your desk, your iMac computer on the left and your iPad on the right make sense together. It would 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. But you would know right away: to you an image is meaningful because it is chock-full of relations that make it what it is and different from countless others.

Therein lies the secret of determining whether a computer is conscious. To do so, pick some images at random from the Web. Black out a strip running vertically down the central third of every one, then shuffle the remaining left and right sides of the pictures. The parts of the composites will not match, except in one case where 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 thwarts the simple image-analysis strategies that computers use today—say, matching lines of texture or color across the separated, partial images. Another test inserts objects into several images so that these objects make sense in all images except one, and the computer must detect the odd one out. A keyboard placed in front of an iMac is the right choice, not a potted plant. A variety of dedicated modules looking for specific high-level features, such as whether a face rests on a neck and so on, might manage to defeat one of these tests. But presenting many different image tests, not unlike

asking many arbitrary questions about the image, would defeat today's machines.

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 data are 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.

In the next Consciousness Redux column, we'll tell you about the surprising results of a near-identical test that psychologists devised to probe the extent to which the unconscious can solve such problems.

This article was originally published with the title "Consciousness Redux: Testing for Consciousness in Machines" in SA Mind 22, 4, 16-17 (September 2011)

doi:10.1038/scientificamericanmind0911-16

(Further Reading)

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

- **A Test for Consciousness. Christof Koch and Giulio Tononi in *Scientific American*, Vol. 304, No. 6, pages 44–47; June 2011.**

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. Koch 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.