Can We Quantify Machine Consciousness?

Artificial intelligence might endow some computers with self-awareness. Here's how we'd know

By **CHRISTOF KOCH AND GIULIO TONONI** Posted 25 May 2017 | 15:00 GMT

Imagine that at some time in the not-too-distant future, you've bought a smartphone that comes bundled with a personal digital assistant (PDA) living in the cloud. You assign a sexy female voice to the PDA and give it access to all of your emails, social media accounts, calendar, photo album, contacts, and other bits and flotsam of your digital life. She—for that's how you quickly think of her—knows you better than your mother, your soon-to-be ex-wife, your friends, or your therapist. Her command of English is flawless; you have endless conversations about daily events; she gets your jokes. She is the last voice you hear before you drift off to sleep and the first upon awakening. You panic when she's off-line. She becomes indispensable to your well-being and so, naturally, you fall in love. Occasionally, you wonder whether she truly reciprocates your feelings and whether she is even capable of experiencing anything at all. But the warm, husky tone of her voice and her ability to be that perfect foil to your narcissistic desires overcome these existential doubts. Alas, your infatuation eventually cools off after you realize she is carrying on equally intimate conversations with thousands of other customers.

This, of course, is the plot of *Her (http://www.imdb.com/title/tt1798709/)*, a 2013 movie in which an anodyne Theodore Twombly falls in love with the software PDA Samantha.

Over the next few decades such a fictional scenario will become real and commonplace. Deep machine learning (https://en.wikipedia.org/wiki/Deep_learning), speech recognition, and related technologies have dramatically progressed, leading to Amazon's Alexa (https://developer.amazon.com/alexa), Apple's Siri (https://www.apple.com/ios/siri/), Google's Now (http://www.androidcentral.com/google-now), and

CAN WE COPY THE BRAIN?

Section 3: Engineering Cognition

(/static/special-report-can-we-copy-the-brain)

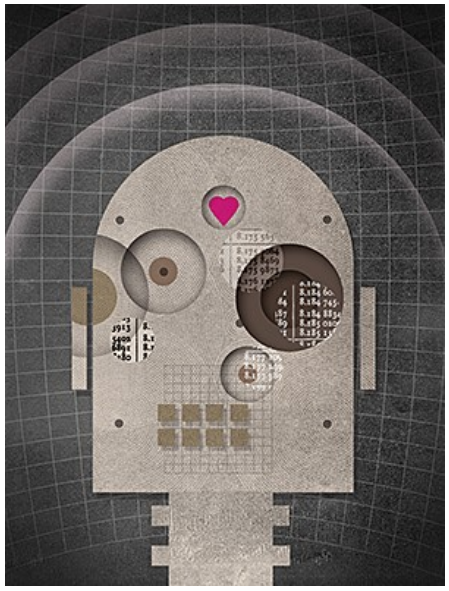


Illustration: Chad Hagen

Microsoft's Cortana (https://www.microsoft.com/en-us/mobile/experiences/cortana/). These virtual assistants will continue to improve until they become hard to distinguish from real people, except that they'll be endowed with perfect recall, poise, and patience—unlike any living being.

The availability of such digital simulacra of many qualities we consider uniquely human will raise profound scientific, psychological, philosophical, and ethical questions. These emulations will ultimately upend the way we think about ourselves, about human exceptionalism, and about our place in the great scheme of things.

Here we will survey the intellectual lay of the land concerning these coming developments. Our view is that as long as such machines are based on present-day computer architectures, they may act just like people—and we may be tempted to treat them that way—but they will, in fact, feel nothing at all. If computers are built more like the brain is, though, they could well achieve true consciousness.

The faith of our age is faith in the digital computer—programmed properly, it will give us all we wish. Cornucopia. Indeed, smart money in Silicon Valley holds that digital computers will be able to replicate and soon exceed anything and everything that humans are capable of.

But could sufficiently advanced computers ever become conscious? One answer comes from those who subscribe to computationalism (https://en.wikipedia.org/wiki/Computational_theory_of_mind), the reigning theory of mind in contemporary philosophy, psychology, and neuroscience. It avers that all mental states—such as your conscious experience of a god-awful toothache or the love you feel for your partner—are computational states. These are fully characterized by their functional relationships to relevant sensory inputs, behavioral outputs, and other computational states in between. That is, brains are elaborate input-output devices that compute and process symbolic representations of the world. Brains are computers, with our minds being the software.

Adherents to computationalism apply these precepts not only to brains and to the behavior they generate but also to the way it *feels* to be a brain in a particular state. After all, that's what consciousness is: any subjective feeling, any experience—what we see, hear, feel, remember, think.

Computationalism assumes that my painful experience of a toothache is but a state of my brain in which certain nerve cells are active in response to the infected tooth, leading to my propensity to moan, hold my jaw, not eat on that side of my mouth, inability to focus on other tasks, and so on. If all of these states are simulated in software on a digital computer, the thinking goes, the system as a whole will not only behave exactly like me but also *feel and think* exactly like me. That is, consciousness is computable. Explicitly or implicitly, this is one of the central tenets held by the digerati in academe, media, and industry.

In this view, there is nothing more to consciousness than the instantiation of the relevant computational states. Nothing else matters, including how the computations are implemented physically, whether on the hardware of a digital computer or on the squishy stuff inside the skull. According to computationalism, a future Samantha—or even better, an embodied example like Ava in the brilliant, dark movie *Ex Machina (http://www.imdb.com/title/tt0470752/)*—will have experiences and feelings just as we do. She will experience sights and sounds, pleasure and pain, love and hate.

Or perhaps she won't.

Computationalism is based on the assumption that if two systems are *functionally* indistinguishable, they will be *mentally* indistinguishable. Because we experience the world, the argument goes, a digital computer that is functionally equivalent to us would necessarily also experience the world as we do—that is, it would also be conscious. But is this assumption warranted? To answer such a question, we need a principled, quantitative theory of what consciousness is and what it takes for a physical system to have it.

Illustration: Chad Hagen

Until recently, such a theory of consciousness wasn't available. True, neuroscientists like us have been engaged in the difficult search for the "neural correlates of consciousness (https://en.wikipedia.org/wiki/Neural_correlates_of_consciousness)," carrying out increasingly elaborate experiments on people and related species such as monkeys and mice. These experiments have identified regions in the neocortex (https://en.wikipedia.org/wiki/Neocortex), the outer surface of the brain just underneath the skull, that are critically involved in consciously seeing and hearing things. Yet having been directly involved in this empirical research program, we know that even if such a quest proves reasonably successful, identifying some particular brain structures or modes of neural activity necessary for consciousness in people or closely related animals will not be sufficient to establish whether creatures with very different nervous systems—such as an octopus or a bee—*are* conscious, by how much, or of what. And any such discovery in neuroscience will be insufficient to establish whether or not machines can be conscious.

There is, however, a fundamental theory of consciousness that offers hope for a principled answer to the question of consciousness in entities vastly different from us, including machines. That theory does not start from behavior or from the brain. Instead, it begins from consciousness itself—from our own experience, the only one we are absolutely certain of. This is the bedrock of certainty that René Descartes, father of modern philosophy, science, and analytic geometry, referred to in the most famous deduction in Western thought: *I think, therefore I am*.

This theory, called integrated information theory, or IIT, has been developed over the past two decades.

It attempts to define what consciousness is, what it takes for a physical system to have it, and how one can measure, at least in principle, both its quantity and its quality, starting from its physical substrate.

IIT is too involved for us to explain here; we can only sketch its general outlines. The theory identifies five essential properties that are true of every conceivable experience of consciousness: (1) Every experience exists intrinsically (for the subject of that experience, not for an external observer); (2) each experience is structured (it is composed of parts and the relations among them); (3) it is integrated (it cannot be subdivided into independent components); (4) it is definite (it has borders, including some contents and excluding others); and (5) it is specific (every experience is the way it is, and thereby different from trillions of possible others).

“The theory can be used to assess the quantity and quality of consciousness for any physical system, whether it is the brain of a human, an octopus, or a bee—or the circuit board of a digital computer”

IIT then translates these properties into requirements that must be satisfied by any physical substrate for it to support consciousness. These requirements can be expressed mathematically and employed to assess the quantity and quality of consciousness for any physical system, whether it is the brain of a human, an octopus, or a bee—or the circuit board of a digital computer.

Crucially, according to IIT, the overall degree of consciousness does not depend on what the system does. Rather, it depends on how it is built—how it’s physically put together. And only certain kinds of physical systems have the right kind of internal architecture to support consciousness: those that have a maximum of intrinsic cause-effect power, the causal power to determine their own states. In essence, this means that the system must be composed of many parts, each having specific causal powers within the overall system (the “information” part of IIT), and yet the system as a whole must not be reducible to those parts (the “integrated” part of IIT), making it far more powerful than the sum of its many parts.

IIT does not use the word “information” in its contemporary sense, as in “messages that are being passed by a sender to a receiver.” Consciousness is not about information sent from one part of the brain to another. Instead, IIT refers to “information” in its original sense, with its root *inform*, meaning “to give form to.” The power of any one mechanism, such as a brain or a computer, to influence its own next state, its causal power, gives rise to a form, a high-dimensional structure, that *is* experience.

IIT can explain in a principled manner many puzzling features of the neuroanatomy of consciousness—for instance, why the cerebellum (https://en.wikipedia.org/wiki/Cerebellum), the little brain underneath the much bigger and better known neocortex, does not contribute to consciousness despite its having four times as many neurons: Its internal architecture, parallel sheets of feed-forward chains of neurons without much recurrent excitation, is very different from the highly heterogeneous, rich, and dense connectivity of the neocortex, which supports vast coalitions of active neurons that quickly assemble and disassemble. It also explains why consciousness fades during certain stages of sleep even though neocortical neurons continue to fire: Parts of the neocortex lose the ability to influence one another effectively.

IIT makes a number of counterintuitive predictions amenable to empirical tests. One prediction is that a nearly silent neocortex, in which few neurons are actively firing, has conscious experiences. Also, IIT has allowed Tononi and Marcello Massimini (https://www.cifar.ca/profiles/marcello-massimini/), now a professor at the University of Milan, to develop a device for assessing consciousness in humans, a combination of a magnetic coil to stimulate the brain and a high-density net of EEG electrodes to detect its response—a crude kind of consciousness meter (http://www.klab.caltech.edu/koch/CR/CR-Consciousness-Meter-13.pdf). This device has already been used to ascertain whether brain-damaged or anesthetized patients unable to communicate are conscious or not.

Being a formal, mathematical theory, IIT can be applied to any physical system, be it the brain—a structure that evolved by natural selection—or an electronic circuit designed by engineers. As ongoing research shows, the physical architecture of certain parts of the neocortex—especially in the back, the way the neurons are connected—is ideal for maximizing the brain's intrinsic cause-effect power, its ability to be affected by its recent state and to determine its future state, which is why it supports consciousness.

By contrast, the physical architecture of a typical digital computer is absolutely inadequate, with very low connectivity at the gate level of its central processing unit and bottlenecks that prevent even a modicum of the necessary integration. That is, a computer may implement computations and functions judged to be intelligent from the perspective of a user looking at its output, but, given its wiring, its intrinsic causal powers as a whole are minute compared with those of any brain. And this is true even if we treat the computer at a coarser level than transistors and resistors.

And here's the rub: That intrinsic power, the physical power to make a difference to oneself, cannot be computed or simulated. It has to be built into the physics of the system. A perfectly executed, biophysically accurate computer simulation of the human brain, including every one of its 86 billion neurons and its matrix of trillions of synapses, wouldn't be conscious. Even if this computer was hooked up to a speech synthesizer and told you about its supposed experiences, it would be nothing but behavior and functions cleverly executing programming. The beating heart of consciousness would be absent.

This consequence of IIT has sobering implications for those who hope that digital brain uploads (http://spectrum.ieee.org/biomedical/imaging/can-machines-be-conscious) may make people immortal. Their vision is that, within the next few decades, we will be able to accurately reconstruct the wiring scheme, the so-called connectome, of any one individual human brain and simulate it on appropriate digital hardware. This process would probably be destructive because there may be no way to access the brain's ultrastructure except by cutting it into wafer-thin slivers. Nevertheless, before you succumb to some deadly disease, you would upload a high-resolution version of your brain to the cloud. As long as the cloud infrastructure is up and running, your digital simulacrum will live on and on, interacting with other digital avatars. Rapture for nerds!

"Although your digital simulacrum might speak and act as you would, it would be a complete

Yet, per IIT, this belief is as illusory as the belief in the hereafter of preceding prophets and religions. Although your digital simulacrum might speak and act as you would, it would be a complete zombie, experiencing nothing. Ironically, though, to your friends and loved ones back in the real world, you would have successfully transitioned to a sublime form of existence and would entice others to join you in an afterlife that, in fact, doesn't exist.

zombie, experiencing nothing"

Whether or not IIT is correct is not merely of academic interest. Barring some global catastrophe, our society will create, within decades, machines with human-level intelligence and behaviors, able to understand speech and talk in many different languages, remember the past and anticipate the future, imagine novel scenarios, write books, compose music, direct films, conceive new goals, as well as move, drive, fly, and, inevitably, fight. From there, thanks to the availability of big data, the power of deep learning (https://en.wikipedia.org/wiki/Deep_learning), and the speed of computing, it will be a short step to overcoming human limits. The birth of true artificial intelligence will profoundly affect mankind's future, including whether it has one.

Whether you are among those who believe that the arrival of human-level AI signals the dawn of paradise or the sunset of the age of humans, you will still have to answer a fundamental question: Are these AIs conscious? Does it *feel* like anything to be them? Or are they immensely more accomplished versions of present-day garbage disposal units, washing machines, or cars—extraordinarily clever machines, yes, but without sentience or feelings?

The answer to this question matters to the way we relate to future machines. If you take a hammer to your shiny Tesla (https://www.tesla.com/), your friends might consider you crazy for destroying such a costly car; yet there is no question that you are free to do so. Try the same with your dog, though, and the police would rightfully arrest you. That's because a car is just a means to an end—a convenient way to get around town—while a dog is an end in itself, with some minimal rights, because it shares with us the gift of consciousness.

"We need a fundamental theory that specifies the exact conditions under which a particular system is capable of conscious experience"

Finding the correct answer, however, cannot be left to our intuition. That might work temporarily for Theodore Twombly falling in love with Samantha, but given the gravity of the situation, we need guidance. We need a fundamental theory that specifies the exact conditions under which a particular system is capable of conscious experience.

IIT predicts that conventional digital computers running software will experience nothing like the movie we see and hear inside our heads. Because smart digital assistants and lifelike future robots are incapable of experience, as IIT insists, their software can be safely copied, edited, sold, pirated, or deleted. And they can be turned off, modified, destroyed, and replaced at will.

But the same need not be true for unconventional architectures. Special-purpose machines built following some of the same design principles as the brain, containing what's called neuromorphic hardware [see "Neuromorphic Chips Are Destined for Deep Learning—or Obscurity," in this issue], could in principle be capable of substantial conscious experience. The key is that the logic and memory gates are heavily interconnected with a high degree of partially overlapping fan-in and fan-out between gates. (Compartmentalized components with highly specific functions do not contribute to intrinsic causal

power.) The way the "brain" of the system is actually wired up, its (bio)physics, makes all the difference, not its input-output behavior.

Such a neuromorphic machine, if highly conscious, would then have intrinsic rights, in particular the right to its own life and well-being. In that case, society would have to learn to share the world with its own creations.

About the Authors

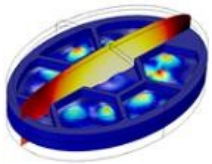
Christof Koch (https://alleninstitute.org/what-we-do/brain-science/about/team/staff-profiles/christof-koch/) is president and chief scientific officer of the Allen Institute for Brain Science, in Seattle. Giulio Tononi (http://centerforsleepandconsciousness.med.wisc.edu/people/tononi.html) holds the David P. White Chair in Sleep Medicine, as well as a Distinguished Chair in Consciousness Science, at the University of Wisconsin.

SPECIAL REPORT: CAN WE COPY THE BRAIN?
(/static/special-report-can-we-copy-the-brain)

PREVIOUS
Why Rat-Brained Robots Are So Good at Navigating Unfamiliar Terrain
(/robotics/robotics-software/why-ratbrained-robots-are-so-good-at-navigating-unfamiliar-terrain)

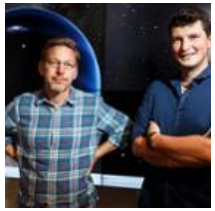
NEXT
Human-Level AI Is Right Around the Corner—or Hundreds of Years Away
(/computing/software/humanlevel-ai-is-right-around-the-corner-or-hundreds-of-years-away)

Recommended For You



Music to Your Ears: New Transducers Meet Electrostatic Headphones

(/computing/software/music-to-your-ears-new-transducers-meet-electrostatic-headphones)



Is There a Giant Planet Lurking Beyond Pluto?

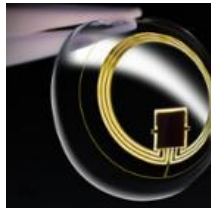
(/aerospace/satellites/is-there-a-giant-planet-lurking-beyond-pluto)

The Big Problem With Self-Driving Cars Is People

(/transportation/self-driving/the-big-problem-with-selfdriving-cars-is-people)

The Transformers: Superheroes of Electrical Inventions

(/energy/the-smarter-grid/the-transformers-superheroes-of-electrical-inventions)



Smart Contact Lenses and Eye Implants Will Give Doctors Medical Insights

(/biomedical/devices/smart-contact-lenses-and-eye-implants-will-give-doctors-medical-insights)

The TurtleBot3 Teacher

(/geek-life/hands-on/the-turtlebot3-teacher)