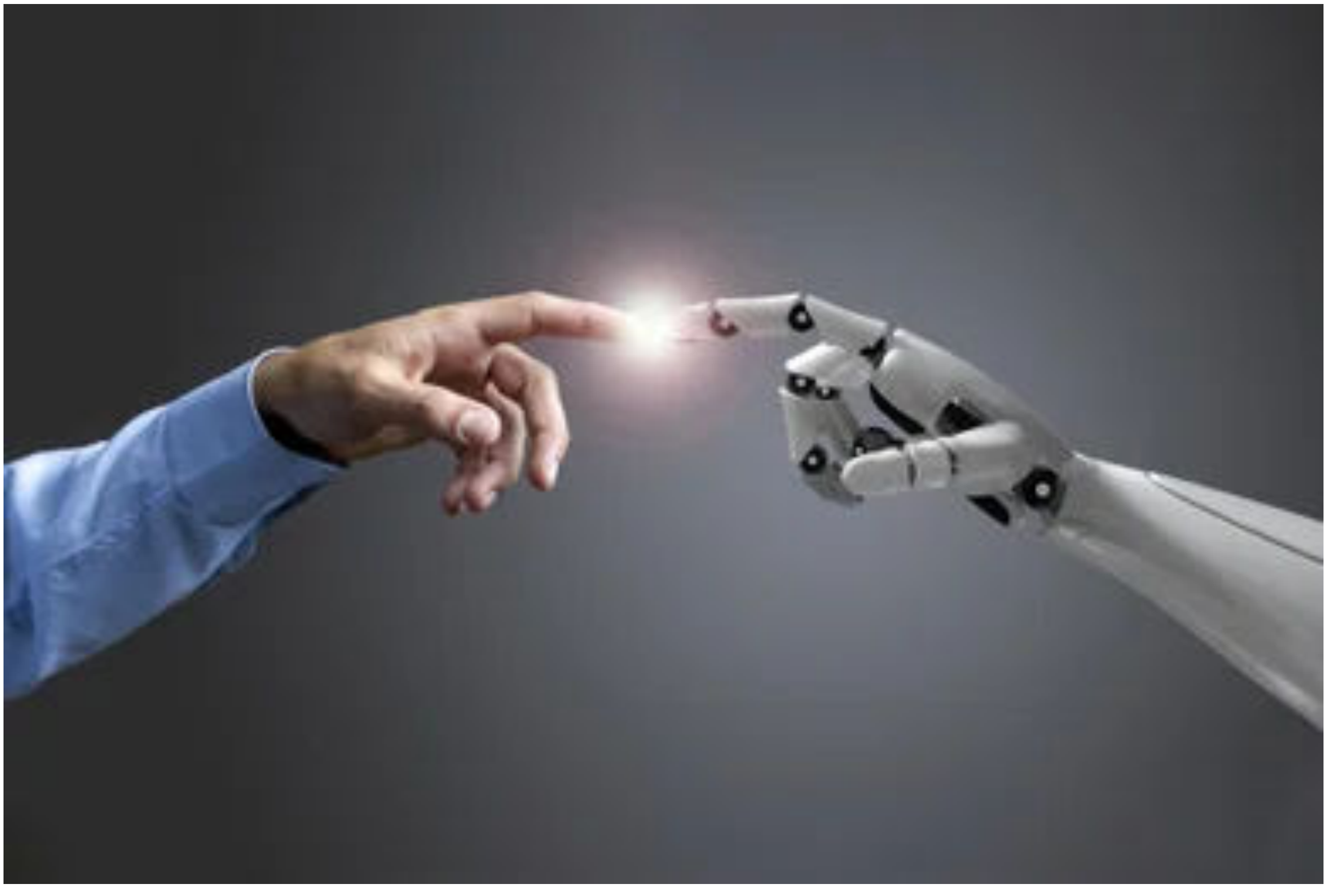
CONSCIOUSNESS

What Does It ‘Feel’ Like to Be a Chatbot?

Generative AI has made giant strides toward machine intelligence. Can machine consciousness be far behind?

By Christof Koch on September 8, 2023



Credit: ClaudioVentrella/Getty Images

The questions of what subjective experience is, who has it and how it relates to the physical world around us have preoccupied philosophers for most of recorded history. Yet the emergence of scientific theories of consciousness that are quantifiable and empirically testable is of much more recent vintage, occurring within the past several decades. Many of these theories focus on the footprints left behind by the subtle cellular networks of the brain from which consciousness emerges.

Progress in tracking these traces of consciousness was very evident at a recent public event in New York City that involved a competition—termed an “adversarial collaboration”—between adherents of today’s two dominant theories of consciousness: integrated information theory (IIT) and global neuronal workspace theory (GNWT). The event came to a head with the resolution of a 25-year-old wager between philosopher of mind David Chalmers of New York University and me.

I had bet Chalmers a case of fine wine that these neural footprints, technically named the neuronal correlates of consciousness, would be unambiguously discovered and described by June 2023. The matchup between IIT and GNWT was left unresolved, given the partially conflicting nature of the evidence concerning which bits and pieces of the brain are responsible for visual experience and the subjective sense of seeing a face or an object, even though the importance of the prefrontal cortex for conscious experiences had been dethroned. Thus, I lost the wager and handed over the wine to Chalmers.

These two dominant theories were developed to explain how the conscious mind relates to neural activity in humans and closely related animals such as monkeys and mice. They make fundamentally different assumptions about subjective experience and come to opposing conclusions with respect to consciousness in engineered artifacts. The extent to which these theories are ultimately empirically verified or falsified for brain-based sentience therefore has important consequences for the looming question of our age: Can machines be sentient?

THE CHATBOTS ARE HERE

Before I come to that, let me provide some context by comparing machines that are conscious with those that display only intelligent behaviors. The holy grail sought by computer engineers is to endow machines with the sort of highly flexible intelligence that enabled *Homo sapiens* to expand out from Africa and eventually populate the entire planet. This is called artificial general intelligence (AGI). Many have argued that AGI is a distant goal. Within the past year, stunning developments in artificial intelligence have taken the world, including experts, by surprise. The advent of eloquent conversational software applications, colloquially called chatbots, transformed the AGI debate from an esoteric topic among science-fiction enthusiasts and Silicon Valley digerati into a debate that conveyed a sense of widespread public malaise about an existential risk to our way of life and to our kind.

These chatbots are powered by large language models, most famously the series of bots called generative pretrained transformers, or GPT, from the company OpenAI in San Francisco. Given the fluidity, literacy and competency of OpenAI’s most recent iteration of these models, GPT-4, it is easy to believe that it has a mind with a personality. Even its odd glitches, known as “hallucinations,” play into this narrative.

GPT-4 and its competitors—Google’s LaMDA and Bard, Meta’s LLaMA and others—are trained on libraries of digitized books and billions of web pages that are publicly accessible via Web crawlers. The genius of a large language model is that it trains itself without supervision by covering up a word or two and trying to predict the missing expression. It does so over and over and over, billions of times, without anyone in the loop. Once the model has learned by ingesting humanity’s collective digital writings, a user prompts it with a sentence or more it has never seen. It will then predict the most likely word, the next after that, and so on. This simple principle led to astounding results in English, German, Chinese, Hindi, Korean and many more tongues including a variety of programming languages.

Tellingly, the foundational essay of AI, which was written in 1950 by British logician Alan Turing under the title “Computing Machinery and Intelligence,” avoided the topic of “can machines think,” which is really another way of asking about machine consciousness. Turing proposed an “imitation game”: Can an observer objectively distinguish between the typed output of a human and a machine when the identity of both are hidden? Today this is known as the Turing test, and chatbots have aced it (even though they cleverly deny that if you ask them directly). Turing’s strategy unleashed decades of relentless advances that led to GPT but elided the problem.

Implicit in this debate is the assumption that artificial intelligence is the same as artificial consciousness, that being smart is the same as being conscious. While intelligence and sentience go together in humans and other evolved organisms, this doesn’t have to be the case. Intelligence is ultimately about reasoning and learning in order to act—learning from one’s own actions and those of other autonomous creatures to better predict and prepare for the future, whether that means the next few seconds (“Uh-oh, that car is heading toward me fast”) or the next few years (“I need to learn how to code”). Intelligence is ultimately about doing.

Consciousness, on the other hand, is about states of being—seeing the blue sky, hearing birds chirp, feeling pain, being in love. For an AI to run amok, it doesn’t matter one iota whether it feels like anything. All that matters is that it has a goal that is not aligned with humanity’s long-term well-being. Whether or not the AI knows what it is trying to do, what would be called self-awareness in humans, is immaterial. The only thing that counts is that it “mindlessly” [*sic*] pursues this goal. So at least conceptually, if we achieved AGI, that would tell us little about whether being such an AGI felt like anything. With this mise-en-scène, let us return to the original question of how a machine might become conscious, starting with the first of the two theories.

IIT starts out by formulating five axiomatic properties of any conceivable subjective experience. The theory then asks what it takes for a neural circuit to instantiate these five properties by switching some neurons on and others off—or alternatively, what it takes for a computer chip to switch some transistors on and others off. The causal interactions within a circuit in a particular state or the fact that two given neurons being active together can turn another neuron on or off, as the case may be, can be unfolded into a high-dimensional causal structure. This structure is identical to the *quality* of the experience, what it feels like, such as why time flows, space feels extended and colors have a particular appearance. This experience also has a *quantity* associated with it, its integrated information. Only a circuit with a maximum of nonzero integrated information exists as a whole and is conscious. The larger the integrated information, the more the circuit is irreducible, the less it can be considered just the superposition of independent subcircuits. IIT stresses the rich nature of human perceptual experiences—just look around to see the lush visual world around you with untold distinctions and relations, or look at a painting by Pieter Brueghel the Elder, a 16th-century Flemish artist who depicted religious subjects and peasant scenes.

The Peasant Wedding is a 1567 or 1568 painting by Flemish Renaissance painter and printmaker Pieter Brueghel the Elder.
Credit: Peter Horree/Alamy Stock Photo

Any system that has the same intrinsic connectivity and causal powers as a human brain will be, in principle, as conscious as a human mind. Such a system cannot be simulated, however, but must be constituted, or built in the image of the brain. Today's digital computers are based on extremely low connectivity (with the output of one transistor wired to the input of a handful of transistors), compared with that of central nervous systems (in which a cortical neuron receives inputs and makes outputs to tens of thousands of other neurons). Thus, current machines, including those that are cloud-based, will not be conscious of anything even though they will be able, in the fullness of time, to do anything that humans can do. In this view, being ChatGPT will never feel like anything. Note this argument has nothing to do with the total number of components, be that neurons or transistors, but the way they are wired up. It is the interconnectivity which determines the overall complexity of the circuit and the number of different configurations it can be in.

The competitor in this contest, GNWT, starts from the psychological insight that the mind is like a theater in which actors perform on a small, lit stage that represents consciousness, with their actions viewed by an audience of processors sitting offstage in the dark. The stage is the central workspace of the mind, with a small working memory capacity for representing a single percept, thought or memory. The various processing modules—vision, hearing, motor control for the eyes, limbs, planning, reasoning, language comprehension and execution —compete for access to this central workspace. The winner displaces the old content, which then becomes unconscious.

The lineage of these ideas can be traced to the blackboard architecture of the early days of AI, so named to evoke the image of people around a blackboard hashing out a problem. In GNWT, the metaphorical stage along with the processing modules were subsequently mapped onto the architecture of the neocortex, the outermost, folded layers of the brain. The workspace is a network of cortical neurons in the front of the brain, with long-range projections to similar neurons distributed all over the neocortex in prefrontal, parietotemporal and cingulate associative cortices. When activity in sensory cortices exceeds a threshold, a global ignition event is triggered across these cortical areas, whereby information is sent to the entire workspace. The act of globally broadcasting this information is what makes it conscious. Data that are not shared in this manner—say, the exact position of eyes or syntactical rules that make up a well-formulated sentence—can influence behavior, but nonconsciously.

From the perspective of GNWT, experience is quite limited, thoughtlike and abstract, akin to the sparse description that might be found in museums, underneath, say, a Brueghel painting: “Indoor scene of peasants, dressed in Renaissance garb, at a wedding, eating and drinking.”

In IIT’s understanding of consciousness, the painter brilliantly renders the phenomenology of the natural world onto a two-dimensional canvas. In GNWT’s view, this apparent richness is an illusion, an apparition, and all that can be objectively said about it is captured in a high-level, terse description.

GNWT fully embraces the mythos of our age, the computer age, that anything is reducible to a computation. Appropriately programmed computer simulations of the brain, with massive feedback and something approximating a central workspace, will consciously experience the world—perhaps not now but soon enough.

IRRECONCILABLE DIFFERENCES

In stark outlines, that's the debate. According to GNWT and other computational functionalist theories (that is, theories that think of consciousness as ultimately a form of computation), consciousness is nothing but a clever set of algorithms running on a Turing machine. It is the functions of the brain that matter for consciousness, not its causal properties. Provided that some advanced version of GPT takes the same input patterns and produces similar output patterns as humans, then all properties associated with us will carry over to the machine, including our most precious possession: subjective experience.

Conversely, for IIT, the beating heart of consciousness is intrinsic causal power, not computation. Causal power is not something intangible or ethereal. It is very concrete, defined operationally by the extent to which the system's past specifies the present state (cause power) and the extent to which the present specifies its future (effect power). And here's the rub: causal power by itself, the ability to make the system do one thing rather than many other alternatives, cannot be simulated. Not now nor in the future. It must be built into the system.

Consider computer code that simulates the field equations of Einstein's general theory of relativity, which relates mass to spacetime curvature. The software accurately models the supermassive black hole located at the center of our galaxy. This black hole exerts such extensive gravitational effects on its surroundings that nothing, not even light, can escape its pull. Thus its name. Yet an astrophysicist simulating the black hole would not get sucked into their laptop by the simulated gravitational field. This seemingly absurd observation emphasizes the difference between the real and the simulated: if the simulation is faithful to reality, spacetime should warp around the laptop, creating a black hole that swallows everything around it.

Of course, gravity is not a computation. Gravity has causal powers, warping the fabric of space-time, and thereby attracting anything with mass. Imitating a black hole's causal powers requires an actual superheavy object, not just computer code. Causal power can't be simulated but must be constituted. The difference between the real and the simulated is their respective causal powers.

That's why it doesn't rain inside a computer simulating a rainstorm. The software is functionally identical to weather yet lacks its causal powers to blow and turn vapor into water drops. Causal power, the ability to make or take a difference to itself, must be built into the system. This is not impossible. A so-called neuromorphic or bionic computer could be as conscious as a human, but that is not the case for the standard von Neumann architecture that is the foundation of all modern computers. Small prototypes of neuromorphic computers have been built in laboratories, such as Intel's second-generation Loihi 2 neuromorphic chip. But a machine with the needed complexity to elicit something resembling human consciousness—or even that of a fruit fly—remains an aspirational wish for the distant future.

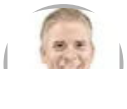
Note that this irreconcilable difference between functionalist and causal theories has nothing to do with intelligence, natural or artificial. As I said above, intelligence is about behaving. Anything that can be produced by human ingenuity, including great novels such as Octavia E. Butler's *Parable of the Sower* or Leo Tolstoy's *War and Peace*, can be mimicked by algorithmic intelligence, provided there is sufficient material to train on. AGI is achievable in the not-too-distant future.

The debate is not about artificial intelligence but about artificial consciousness. This debate cannot be resolved by building bigger language models or better neural network algorithms. The question will need to be answered by understanding the only subjectivity we are indubitably confident of: our own. Once we have a solid explanation of human consciousness and its neural underpinnings, we can extend such an understanding to intelligent machines in a coherent and scientifically satisfactory manner.

The debate matters little to how chatbots will be perceived by society at large. Their linguistic skills, knowledge base and social graces will soon become flawless, endowed with perfect recall, competence, poise, reasoning abilities and intelligence. Some even proclaim that these creatures of big tech are the next step in evolution, Friedrich Nietzsche's "*Übermensch*." I take a darker viewer and believe that these folks mistake our species' dusk for its dawn.

For many, and perhaps for most people in an increasingly atomized society that is removed from nature and organized around social media, these agents, living in their phones, will become emotionally irresistible. People will act, in ways both small and large, like these chatbots are conscious, like they can truly love, be hurt, hope and fear, even if they are nothing more than sophisticated lookup tables. They will become indispensable to us, perhaps more so than truly sentient organisms, even though they feel as much as a digital TV or toaster—nothing.

ABOUT THE AUTHOR(S)



Christof Koch is meritorious investigator at the Allen Institute in Seattle and the chief scientist at the Tiny

Blue Dot Foundation in Santa Monica, as well as author of the forthcoming *Then I am Myself the World – What Consciousness is and How to Expand It* . He serves on Scientific American's board of advisers.
Credit: Nick Higgins

Recent Articles by Christof Koch

A Cell Atlas Reveals the Biodiversity inside Our Head

Electrodes That Stimulate the Brain Reveal the Roots of Conscious Experience

What Near-Death Experiences Reveal about the Brain

Scientific American is part of Springer Nature, which owns or has commercial relations with thousands of scientific publications (many of them can be found at www.springernature.com/us). Scientific American maintains a strict policy of editorial independence in reporting developments in science to our readers.