Title
What is consciousness, and could machines have it?

Journal
Science, 358(6362)

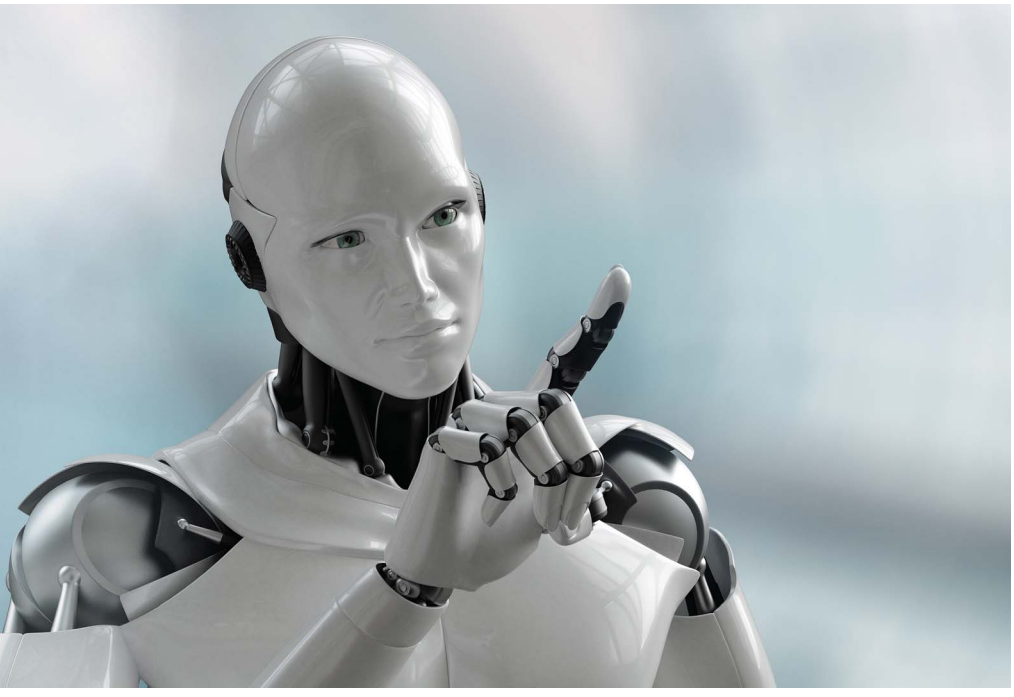
ISSN
0036-8075

Authors
Dehaene, Stanislas
Lau, Hakwan
Kouider, Sid

Publication Date
2017-10-27

DOI
10.1126/science.aan8871

Peer reviewed



CREDIT: MENNOVANDIJK/ISTOCK

REVIEW

What is consciousness, and could machines have it?

Stanislas Dehaene,[1,2]* Hakwan Lau,[3,4] Sid Kouider[5]

The controversial question of whether machines may ever be conscious must be based on a careful consideration of how consciousness arises in the only physical system that undoubtedly possesses it: the human brain. We suggest that the word "consciousness" conflates two different types of information-processing computations in the brain: the selection of information for global broadcasting, thus making it flexibly available for computation and report (C1, consciousness in the first sense), and the self-monitoring of those computations, leading to a subjective sense of certainty or error (C2, consciousness in the second sense). We argue that despite their recent successes, current machines are still mostly implementing computations that reflect unconscious processing (C0) in the human brain. We review the psychological and neural science of unconscious (C0) and conscious computations (C1 and C2) and outline how they may inspire novel machine architectures.

Conscious robots and computers has long been a popular science fiction theme. Why do they lack consciousness in reality, and how might they develop it?

Imagine that you are driving when you suddenly realize that the fuel-tank light is on. What makes you, a complex assembly of neurons, aware of the light? And what makes the car, a sophisticated piece of electronics and engineering, unaware of it? What would it take for the car to be endowed with a consciousness similar to our own? Are those questions scientifically tractable?

Alan Turing and John von Neumann, the founders of the modern science of computation, entertained the possibility that machines would ultimately mimic all of the brain's abilities, including consciousness. Recent advances in artificial intelligence (AI) have revived this goal. Refinements in machine learning, inspired by neurobiology, have led to artificial neural networks that approach or, occasionally, surpass humans (*1*, *2*). Although those networks do not mimic the biophysical properties of actual brains, their design benefitted from several neurobiological insights, including nonlinear input-output functions, layers with converging projections, and modifiable synaptic weights. Advances in computer hardware and training algorithms now allow such networks to operate on complex problems (such as machine translation) with success rates previously thought to be the privilege of real brains. Are they on the verge of consciousness?

We argue that the answer is negative: The computations implemented by current deep-learning networks correspond mostly to nonconscious operations in the human brain. However, much like artificial neural networks took their inspiration from neurobiology, artificial consciousness may progress by investigating the architectures that allow the human brain to generate consciousness, then transferring those insights into computer algorithms. Our aim is to foster such progress by reviewing aspects of the cognitive neuroscience of consciousness that may be pertinent for machines.

Multiple meanings of consciousness

The word "consciousness," like many prescientific terms, is used in widely different senses. In a medical context, it is often used in an intransitive sense (as in, "the patient was no longer conscious"), in the context of assessing vigilance and wakefulness. Elucidating the brain mechanisms of vigilance is an essential scientific goal, with major consequences for our understanding of sleep, anesthesia, coma, or vegetative state. For lack of space, we do not deal with this aspect here, however, because its computational impact seems minimal: Obviously, a machine must be properly turned on for its computations to unfold normally.

We suggest that it is useful to distinguish two other essential dimensions of conscious computation. We label them using the terms global availability (C1) and self-monitoring (C2).

C1: Global availability

This corresponds to the transitive meaning of consciousness (as in "The driver is conscious of the light"). It refers to the relationship between a cognitive system and a specific object of thought, such as a mental representation of "the fuel-tank light." This object appears to be selected for further processing, including verbal and nonverbal report. Information that is conscious in this sense becomes globally available to the organism; for example, we can recall it, act upon it, and speak about it. This sense is synonymous with "having the information in mind"; among the vast repertoire of thoughts that can become conscious at a given time, only that which is globally available constitutes the content of C1 consciousness.

C2: Self-monitoring

Another meaning of consciousness is reflexive. It refers to a self-referential relationship in which

[1]Chair of Experimental Cognitive Psychology, Collège de France, 11 Place Marcelin Berthelot, 75005 Paris, France. [2]Cognitive Neuroimaging Unit, Commissariat à l'Energie Atomique et aux Energies Alternatives (CEA), INSERM, Université Paris-Sud, Université Paris-Saclay, NeuroSpin center, 91191 Gif/Yvette, France. [3]Department of Psychology and Brain Research Institute, University of California, Los Angeles, Los Angeles, CA, USA. [4]Department of Psychology, University of Hong Kong, Pokfulam Road, Hong Kong. [5]Brain and Consciousness Group (École Normale Supérieure, École des Hautes Études en Sciences Sociales, CNRS), Département d'Études Cognitives, École Normale Supérieure–Paris Sciences et Lettres Research University, Paris, France.
*Corresponding author. Email: stanislas.dehaene@cea.fr

the cognitive system is able to monitor its own processing and obtain information about itself. Human beings know a lot about themselves, including such diverse information as the layout and position of their body, whether they know or perceive something, or whether they just made an error. This sense of consciousness corresponds to what is commonly called introspection, or what psychologists call "meta-cognition"—the ability to conceive and make use of internal representations of one's own knowledge and abilities.

We propose that C1 and C2 constitute orthogonal dimensions of conscious computations. This is not to say that C1 and C2 do not involve overlapping physical substrates; in fact, as we review below, in the human brain both depend on the prefrontal cortex. But we argue that empirically and conceptually, the two may come apart because there can be C1 without C2, such as when reportable processing is not accompanied by accurate metacognition, or C2 without C1, such as when a self-monitoring operation unfolds without being consciously reportable. As such, it is advantageous to consider these computations separately before we consider their synergy. Furthermore, many computations involve neither C1 nor C2 and therefore are properly called "unconscious" (or C0 for short). It was Turing's original insight that even sophisticated information processing can be realized by a mindless automaton. Cognitive neuroscience confirms that complex computations such as face or speech recognition, chess-game evaluation, sentence parsing, and meaning extraction occur unconsciously in the human brain—under conditions that yield neither global reportability nor self-monitoring (Table 1). The brain appears to operate, in part, as a juxtaposition of specialized processors or "modules" that operate nonconsciously and, we argue, correspond tightly to the operation of current feedforward deep-learning networks.

We next review the experimental evidence for how human and animal brains handle C0-, C1-, and C2-level computations, before returning to machines and how they could benefit from this understanding of brain architecture.

Unconscious processing (C0): Where most of our intelligence lies

"We cannot be conscious of what we are not conscious of" (*3*). This truism has deep consequences. Because we are blind to our unconscious processes, we tend to underestimate their role in our mental life. However, cognitive neuroscientists developed various means of presenting images or sounds without inducing any conscious experience (Fig. 1) and then used behavioral and brain imaging to probe their processing depth.

The phenomenon of priming illustrates the remarkable depth of unconscious processing. A highly visible target stimulus, such as the written word "four," is processed more efficiently when preceded by a related prime stimulus, such as the Arabic digit "4," even when subjects do not notice the presence of the prime and cannot reliably report its identity. Subliminal digits, words, faces, or objects can be invariantly recognized and influence motor, semantic, and decision levels of processing (Table 1). Neuroimaging methods reveal that the vast majority of brain areas can be activated nonconsciously.

Unconscious view-invariance and meaning extraction in the human brain

Many of the difficult perceptual computations, such as invariant face recognition or speaker-invariant speech recognition, that were recently addressed by AI correspond to nonconscious computations in the human brain (*4*–*6*). For instance, processing someone's face is facilitated when it is preceded by the subliminal presentation of a totally different view of the same person, indicating unconscious invariant recognition (Fig. 1). Subliminal priming generalizes across visual-auditory modalities (*7*, *8*), revealing that cross-modal computations that remain challenging for AI software (such as extraction of semantic vectors or speech-to-text) also involve unconscious mechanisms. Even the semantic meaning of sensory input can be processed without awareness by the human brain. Compared with related words (for example, animal-dog), semantic violations (for example, furniture-dog) generate a brain response as late as 400 ms after stimulus onset in temporal-lobe language networks, even if one of the two words cannot be consciously detected (*9*, *10*).

Unconscious control and decision-making

Unconscious processes can reach even deeper levels of the cortical hierarchy. For instance, subliminal primes can influence prefrontal mechanisms of cognitive control involved in the selection of a task (*11*) or the inhibition of a motor response (*12*). Neural mechanisms of decision-making involve accumulating sensory evidence that affects the probability of the various choices until a threshold is attained. This accumulation of probabilistic knowledge continues to happen even with subliminal stimuli (*13*–*16*). Bayesian inference and evidence accumulation, which are cornerstone computations for AI (*2*), are basic unconscious mechanisms for humans.

Unconscious learning

Reinforcement learning algorithms, which capture how humans and animals shape their future

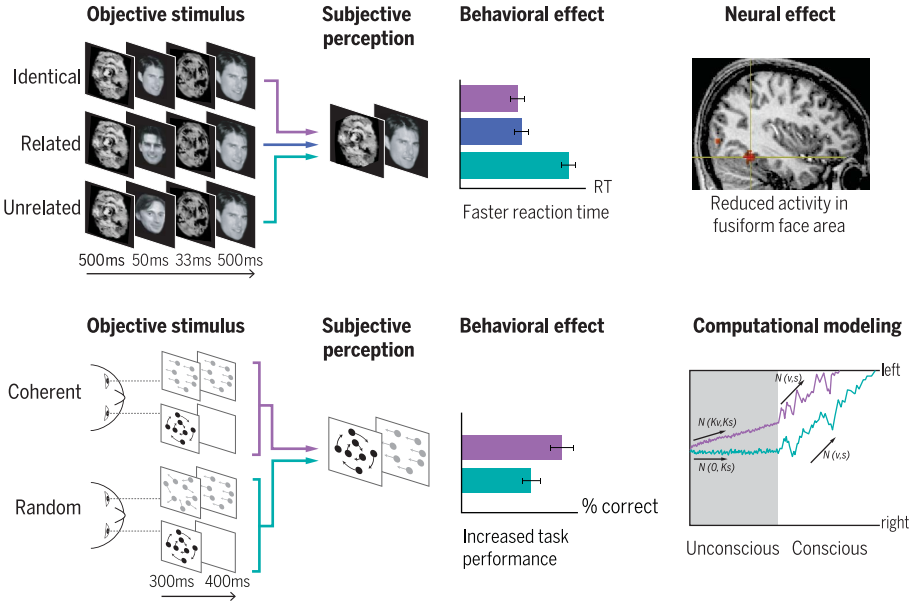


Fig. 1. Examples of paradigms probing unconscious processing (C0). (**Top**) Subliminal view-invariant face recognition (*77*). On each trial, a prime face is briefly presented (50 ms), surrounded by masks that make it invisible, followed by a visible target face (500 ms). Although subjective perception is identical across conditions, processing is facilitated whenever the two faces represent the same person, in same or different view. At the behavioral level, this view-invariant unconscious priming is reflected in reduced reaction time in recognizing the target face. At the neural level, it is reflected in reduced cortical response to the target face (repetition suppression) in the fusiform face area of the human inferotemporal cortex. (**Bottom**) Subliminal accumulation of evidence during interocular suppression (*16*). Presentation of salient moving dots in one eye prevents the conscious perception of paler moving dots in the opposite eye. Despite their invisibility, the gray dots facilitate performance when they moved in the same direction as a subsequent dot display, an effect proportional to their amount of motion coherence. This facilitation only affects a first-order task (judging the direction of motion), not a second-order metacognitive judgement (rating the confidence in the first response). A computational model of evidence accumulation proposes that subliminal motion information gets added to conscious information, thus biasing and shortening the decision.

actions on the basis of history of past rewards, have excelled in attaining supra-human AI performance in several applications, such as playing Go (*1*). Remarkably, in humans, such learning appears to proceed even when the cues, reward, or motivation signals are presented below the consciousness threshold (*17*, *18*).

Complex unconscious computations and inferences routinely occur in parallel within various brain areas. Many of these C0 computations have now been captured by AI, particularly by using feedforward convolutional neural networks (CNNs). We next consider what additional computations are required for conscious processing.

C1: Global availability of relevant information

The need for integration and coordination

The organization of the brain into computationally specialized subsystems is efficient, but this architecture also raises a specific computational problem: The organism as a whole cannot stick to a diversity of probabilistic interpretations; it must act and therefore cut through the multiple possibilities and decide in favor of a single course of action. Integrating all of the available evidence to converge toward a single decision is a computational requirement that, we contend, must be faced by any animal or autonomous AI system and corresponds to our first functional definition of consciousness: global availability (C1).

For example, elephants, when thirsty, manage to determine the location of the nearest water hole and move straight to it, from a distance of 5 to 50 km (*19*). Such decision-making requires a sophisticated architecture for (i) efficiently pooling over all available sources of information, including multisensory and memory cues; (ii) considering the available options and selecting the best one on the basis of this large information pool; (iii) sticking to this choice over time; and (iv) coordinating all internal and external processes toward the achievement of that goal. Primitive organisms, such as bacteria, may achieve such decision solely through an unconscious competition of uncoordinated sensorimotor systems. This solution, however, fails as soon as it becomes necessary to bridge over temporal delays and to inhibit short-term tendencies in favor of longer-term winning strategies. Coherent, thoughtful planning required a specific C1 architecture.

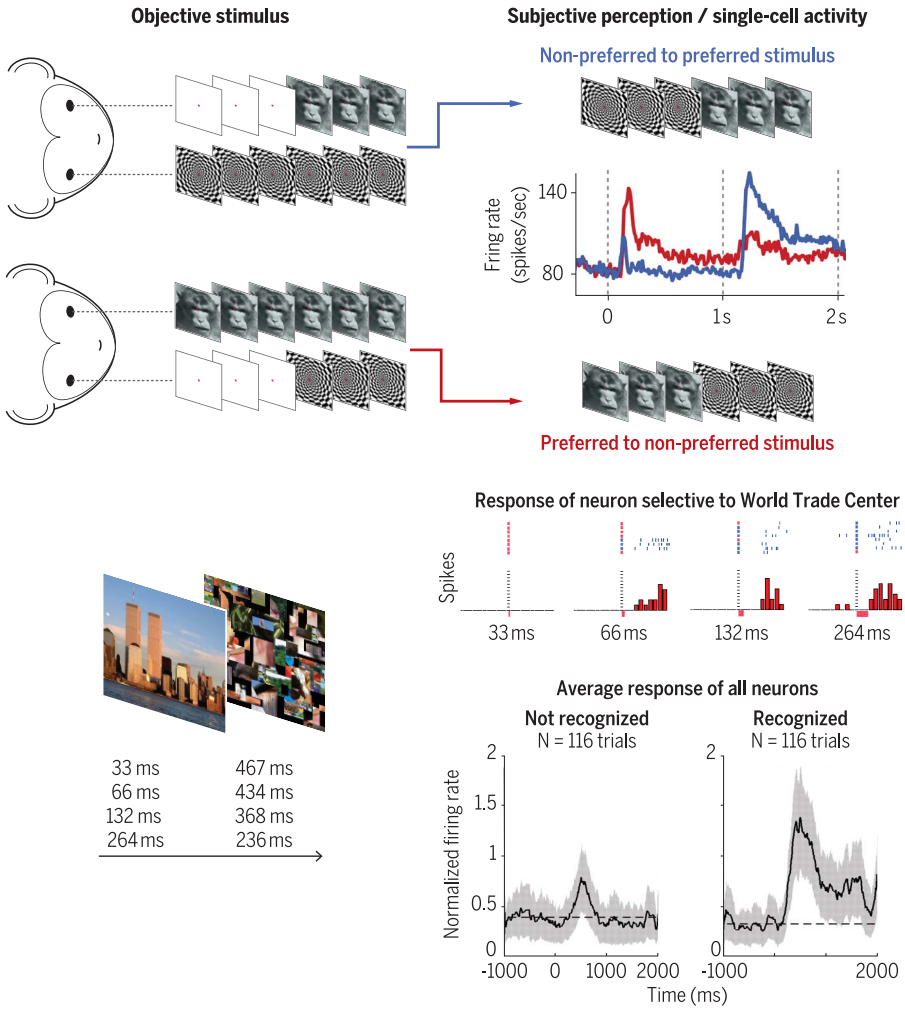


Fig. 2. Global availability: Consciousness in the first sense (C1). Conscious subjective percepts are encoded by the sudden firing of stimulus-specific neural populations distributed in interconnected, high-level cortical areas, including the lateral prefrontal cortex, anterior temporal cortex, and hippocampus. (**Top**) During binocular flash suppression, the flashing of a picture to one eye suppresses the conscious perception of a second picture presented to the other eye. As a result, the same physical stimulus can lead to distinct subjective percepts. This example illustrates a prefrontal neuron sensitive to faces and unresponsive to checkers, whose firing shoots up in tight association with the sudden onset of subjective face perception (*31*). (**Bottom**) During masking, a flashed image, if brief enough and followed by a longer "mask," can remain subjectively invisible. Shown is a neuron in the entorhinal cortex firing selectively to the concept of "World Trade Center." Rasters in red indicate trials in which the subject reported recognizing the picture (blue indicates no recognition). Under masking, when the picture is presented for only 33 ms there is little or no neural activity; but once presentation time is longer than the perceptual threshold (66 ms or larger), the neuron fires substantially only on recognized trials. Overall, even for identical objective input (same duration), spiking activity is higher and more stable for recognized trials (*38*).

Consciousness as access to an internal global workspace

We hypothesize that consciousness in the first sense (C1) evolved as an information-processing architecture that addresses this information-pooling problem (*20*–*23*). In this view, the architecture of C1 evolved to break the modularity and parallelism of unconscious computations. On top of a deep hierarchy of specialized modules, a "global neuronal workspace," with limited capacity, evolved to select a piece of information, hold it over time, and share it across modules. We call "conscious" whichever representation, at a given time, wins the competition for access to this mental arena and gets selected for global sharing and decision-making. Consciousness is therefore manifested by the temporary dominance of a thought or train of thoughts over mental processes, so that it can guide a broad variety of behaviors. These behaviors include not only physical actions but also mental ones, such as committing information to episodic memory or routing it to other processors.

Relation between consciousness and attention

William James described attention as "the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought" (*24*). This definition is close to what we mean by C1: the selection of a single piece of information for entry into the global workspace. There is, however, a clear-cut distinction between this final step, which corresponds to conscious access, and the previous stages of attentional selection, which can operate unconsciously. Many experiments have established

Table 1. Examples of computations pertaining to information-processing levels C0, C1 and C2 in the human brain.

Computation	Examples of experimental findings	References
	C0: Unconscious processing	
Invariant visual recognition	Subliminal priming by unseen words and faces, invariant for font, size, or viewpoint.	(*5*)
	Functional MRI (fMRI) and single-neuron response to unseen words and faces	(*33*, *37*, *78*, *79*)
	Unconscious judgement of chess game configurations	(*80*)
Access to meaning	N400 response to unseen out-of-context words	(*9*, *10*)
Cognitive control	Unconscious inhibition or task set preparation by an unseen cue	(*11*, *12*)
Reinforcement learning	Subliminal instrumental conditioning by unseen shapes	(*17*)
	C1: Global availability of information	
All-or-none selection and broadcasting of a relevant content	Conscious perception of a single picture during visual rivalry	(*29*)
	Conscious perception of a single detail in a picture or stream	(*28*, *81*)
	All-or-none memory retrieval	(*82*)
	Attentional blink: Conscious perception of item A prevents the simultaneous perception of item B	(*27*, *30*, *83*, *84*)
	All-or-none "ignition" of event-related potentials and fMRI signals, only on trials with conscious perception	(*33–35*, *85–87*)
	All-or-none firing of neurons coding for the perceived object in prefrontal cortex and other higher areas	(*31*, *32*, *37*, *38*, *88*)
Stabilization of short-lived information for off-line processing	Brain states are more stable when information is consciously perceived; unconscious information quickly decays (~1 s)	(*39*, *89*)
	Conscious access may occur long after the stimulus is gone	(*90*)
Flexible routing of information	Only conscious information can be routed through a series of successive operations (for example, successive calculations 3 × 4 + 2)	(*91*)
Sequential performance of several tasks	Psychological refractory period: Conscious processing of item A delays conscious processing of item B	(*34*, *92*)
	Serial calculations or strategies require conscious perception	(*13*, *91*)
	Serial organization of spontaneous brain activity during conscious thought in the "resting state"	(*93*)
	C2: Self-monitoring	
Self-confidence	Humans accurately report subjective confidence, a probabilistic estimate in the accuracy of a decision or computation	(*51*, *55*)
Evaluation of one's knowledge	Humans and animals can ask for help or "opt out" when unsure	(*53*, *65*, *66*)
	Humans and animals know when they do not know or remember	(*49*, *53*)
Error detection	Anterior cingulate response to self-detected errors	(*61*, *65*, *94*)
Listing one's skills	Children know the arithmetic procedures at their disposal, their speed, and error rate.	(*70*)
Sharing one's confidence with others	Decision-making improves when two persons share knowledge	(*69*)

the existence of dedicated mechanisms of attention orienting and shown that, like any other processors, they can operate nonconsciously: (i) In the top-down direction, attention can be oriented toward an object, amplify its processing, and yet fail to bring it to consciousness (*25*); (ii) in the bottom-up direction, attention can be attracted by a flash, even if this stimulus ultimately remains unconscious (*26*). What we call attention is a hierarchical system of sieves that operate unconsciously. Such unconscious systems compute with probability distributions, but only a single sample, drawn from this probabilistic distribution, becomes conscious at a given time (*27*, *28*). We may become aware of several alternative interpretations, but only by sampling their unconscious distributions over time (*29*, *30*).

Evidence for all-or-none selection in a capacity-limited system

The primate brain comprises a conscious bottleneck and can only consciously access a single item at a time (Table 1). For instance, rivaling pictures or ambiguous words are perceived in an all-or-none manner; at any given time, we subjectively perceive only a single interpretation out of many possible ones [even though the others continue to be processed unconsciously (*31*, *32*)]. The serial operation of consciousness is attested by phenomena such as the attentional blink and the psychological refractory period, in which conscious access to a first item A prevents or delays the perception of a second competing item B (*9*, *27*, *30*, *33–35*). Such interference with the perception of B is triggered by the mere conscious perception of A, even if no task is performed (*36*). Thus C1 consciousness is causally responsible for a serial information-processing bottleneck.

Evidence for integration and broadcasting

Brain imaging in humans and neuronal recordings in monkeys indicate that the conscious bottleneck is implemented by a network of neurons that is distributed through the cortex, but with a stronger emphasis on high-level associative areas. Listed in Table 1 are some of the publications that have evidenced an all-or-none "ignition" of this network during conscious perception by using a variety of brain-imaging techniques. Single-cell recordings indicate that each specific conscious percept, such as a person's face, is encoded by the all-or-none firing of a subset of neurons in high-level temporal and prefrontal cortices, whereas others remain silent (Fig. 2) (*31*, *32*, *37*, *38*).

Stability as a feature of consciousness

Direct contrasts between seen and unseen pictures or words confirm that such ignition occurs only for the conscious percept. As explained earlier, nonconscious stimuli may reach into deep cortical networks and influence higher levels of processing and even central executive functions, but these effects tend to be small, variable, and short-lived [although nonconscious information decays at a slower rate than initially expected (*39*, *40*)]. By contrast, the stable, reproducible representation of high-quality information by a distributed activity pattern in higher cortical areas is a feature of conscious processing (Table 1). Such transient

"meta-stability" seems to be necessary for the nervous system to integrate information from a variety of modules and then broadcast it back to them, achieving flexible cross-module routing.

C1 consciousness in human and nonhuman animals

C1 consciousness is an elementary property that is present in human infants (*41*) as well as in animals. Nonhuman primates exhibit similar visual illusions (*31*, *32*), attentional blink (*42*), and central capacity limits (*43*) as human subjects. The prefrontal cortex appears to act as a central information sharing device and serial bottleneck in both human and nonhuman primates (*43*). The considerable expansion of the prefrontal cortex in the human lineage may have resulted in a greater capacity for multimodal convergence and integration (*44*–*46*). Furthermore, humans possess additional circuits in the inferior prefrontal cortex for verbally formulating and reporting information to others. The capacity to report information through language is universally considered one of the clearest signs of conscious perception because once information has reached this level of representation in humans, it is necessarily available for sharing across mental modules and therefore conscious in the C1 sense. Thus, although language is not required for conscious perception and processing, the emergence of language circuits in humans may have resulted in a considerable increase in the speed, ease, and flexibility of C1-level information sharing.

C2: Self-monitoring

Whereas C1 consciousness reflects the capacity to access external information, consciousness in the second sense (C2) is characterized by the ability to reflexively represent oneself (*47*–*50*). A substantial amount of research in cognitive neuroscience and psychology has addressed self-monitoring under the term of "metacognition," which is roughly defined as cognition about cognition or knowing about knowing. Below, we review the mechanisms by which the primate brain monitors itself, while stressing their implications for building self-reflective machines.

A probabilistic sense of confidence

When making a decision, humans feel more or less confident about their choice. Confidence can be defined as a sense of the probability that a decision or computation is correct (*51*). Almost anytime the brain perceives or decides, it also estimates its degree of confidence. Learning is also accompanied by a quantitative sense of confidence; humans evaluate how much trust they have in what they have learned and use it to weigh past knowledge versus present evidence (*52*). Confidence can be assessed nonverbally, either retrospectively, by measuring whether humans persist in their initial choice, or prospectively, by allowing them to opt out from a task without even attempting it. Both measures have been used in nonhuman animals to show that they too possess metacognitive abilities (*53*). By contrast, most current neural networks lack them: Although they can learn, they generally lack meta-knowledge of the reliability and limits of what has been learned. A noticeable exception is biologically constrained models that rely on Bayesian mechanisms to simulate the integration of multiple probabilistic cues in neural circuits (*54*). These models have been fruitful in describing how neural populations may automatically compute the probability that a given process is performed successfully. Although these implementations remain rare and have not addressed the same range of computational problems as has traditional AI, they offer a promising venue for incorporating uncertainty monitoring in deep learning networks.

Explicit confidence in prefrontal cortex

According to Bayesian accounts, each local cortical circuit may represent and combine probability distributions in order to estimate processing uncertainty (*54*). However, additional neural circuits may be required in order to explicitly extract and manipulate confidence signals. Magnetic resonance imaging (MRI) studies in humans and physiological recordings in primates and even in rats have specifically linked such confidence processing to the prefrontal cortex (*55*–*57*). Inactivation of the prefrontal cortex can induce a specific deficit in second-order (metacognitive) judgements while sparing performance on the first-order task (*56*, *58*). Thus, circuits in the prefrontal cortex may have evolved to monitor the performance of other brain processes.

> ***"...it is useful to distinguish two other essential dimensions of conscious computation."***

Error detection: Reflecting on one's own mistakes

Error detection provides a particularly clear example of self-monitoring; just after responding, we sometimes realize that we made an error and change our mind. Error detection is reflected by two components of electroencephalography (EEG) activity: the error-relativity negativity (ERN) and the positivity upon error (Pe), which emerge in cingulate and medial prefrontal cortex just after a wrong response but before any feedback is received. How can the brain make a mistake and detect it? One possibility is that the accumulation of sensory evidence continues after a decision is made, and an error is inferred whenever this further evidence points in the opposite direction (*59*). A second possibility, more compatible with the remarkable speed of error detection, is that two parallel circuits, a low-level sensory-motor circuit and a higher-level intention circuit, operate on the same sensory data and signal an error whenever their conclusions diverge (*60*, *61*).

Meta-memory

Humans do not just know things about the world; they actually know that they know or that they do not know. A familiar example is having a word "on the tip of the tongue." The term "meta-memory" was coined to capture the fact that humans report feelings of knowing, confidence, and doubts on their memories. Meta-memory is thought to involve a second-order system that monitors internal signals (such as the strength and quality of a memory trace) to regulate behavior. Meta-memory is associated with prefrontal structures whose pharmacological inactivation leads to a metacognitive impairment while sparing memory performance itself (*56*). Metamemory is crucial to human learning and education by allowing learners to develop strategies such as increasing the amount of study or adapting the time allocated to memory encoding and rehearsal (*49*).

Reality monitoring

In addition to monitoring the quality of sensory and memory representations, the human brain must also distinguish self-generated versus externally driven representations. Indeed, we can perceive things, but we also conjure them from imagination or memory. Hallucinations in schizophrenia have been linked to a failure to distinguish whether sensory activity is generated by oneself or by the external world (*62*). Neuroimaging studies have linked this kind of reality monitoring to the anterior prefrontal cortex (*63*). In nonhuman primates, neurons in the prefrontal cortex distinguish between normal visual perception and active maintenance of the same visual content in memory (*64*).

Foundations of C2 consciousness in infants

Self-monitoring is such a basic ability that it is already present during infancy (Fig. 3). The ERN, indicating error monitoring, was observed when 1-year-old infants made a wrong choice in a perceptual decision task (*65*). Similarly, after 1-½-year-old infants pointed to one of two boxes in order to obtain a hidden toy, they waited longer for an upcoming reward (such as a toy) when their initial choice was correct than when it was wrong, suggesting that they monitored the likelihood that their decision was right (*57*, *65*). Moreover, when given the opportunity to ask (nonverbally) their parents for help they chose this opt-out option specifically in trials in which they were likely to be wrong, revealing a prospective estimate of their own uncertainty (*66*). That infants can communicate their own uncertainty to other agents further suggests that they consciously experience metacognitive information. Thus, infants are already equipped with the ability to monitor their own mental states. Facing a world where everything remains to be learned, C2 mechanisms allow them to actively orient toward domains that they know they do not know—a mechanism that we call "curiosity."

Dissociations between C1 and C2

According to our analysis, C1 and C2 are largely orthogonal and complementary dimensions of what we call consciousness. On one side of this double dissociation, self-monitoring can exist for

unreportable stimuli (C2 without C1). Automatic typing provides a good example: Subjects slow down after a typing mistake, even when they fail to consciously notice the error (*67*). Similarly, at the neural level, an ERN can occur for subjectively undetected errors (*68*). On the other side of this dissociation, consciously reportable contents sometimes fail to be accompanied with an adequate sense of confidence (C1 without C2). For instance, when we retrieve a memory, it pops into consciousness (C1) but sometimes without any accurate evaluation of its confidence (C2), leading to false memories. As noted by Marvin Minsky, "what we call consciousness [in the C1 sense] is a very imperfect summary in one part of the brain of what the rest is doing." The imperfection arises in part from the fact that the global workspace reduces complex parallel sensory streams of probabilistic computation to a single conscious sample (*27*–*29*). Thus, probabilistic information is often lost on the way, and subjects feel over-confident in the accuracy of their perception.

Synergies between C1 and C2 consciousness

Because C1 and C2 are orthogonal, their joint possession may have synergistic benefits to organisms. In one direction, bringing probabilistic metacognitive information (C2) into the global workspace (C1) allows it to be held over time, integrated into explicit long-term reflection, and shared with others. Social information sharing improves decisions: By sharing their confidence signals, two persons achieve a better performance in collective decision-making than that of either person alone (*69*). In the converse direction, the possession of an explicit repertoire of one's own abilities (C2) improves the efficiency with which C1 information is processed. During mental arithmetic, children can perform a C2-level evaluation of their available competences (for example, counting, adding, multiplying, or memory retrieval) and use this information to evaluate how to best face a given arithmetic problem (*70*). This functionality requires a single "common currency" for confidence across different modules, which humans appear to possess (*71*).

Endowing machines with C1 and C2

How could machines be endowed with C1 and C2 computations? Let us return to the car light example. In current machines, the "low gas" light is a prototypical example of an unconscious modular signal (C0). When the light flashes, all other processors in the machine remain uninformed and unchanged; fuel continues to be injected in the carburetor, and the car passes gas stations without stopping (although they might be present on the GPS map). Current cars or cell phones are mere collections of specialized modules that are largely "unaware" of each other. Endowing this machine with global information availability (C1) would allow these modules to share information and collaborate to address the impending problem (much like humans do when they become aware of the light, or elephants of thirst).

Although AI has met considerable success in solving specific problems, implementing multiple processes in a single system and flexibly coordinating them remain difficult problems. In the 1960s, computational architectures called "blackboard systems" were specifically designed to post information and make it available to other modules in a flexible and interpretable manner, similar in flavor to a global workspace (*20*). A recent architecture called Pathnet uses a genetic algorithm to learn which path through its many specialized neural networks is most suited to a given task (*72*). This architecture exhibits robust, flexible performance and generalization across tasks and may constitute a first step toward primate-like conscious flexibility.

"...the ability to reflexively represent oneself."

To make optimal use of the information provided by the fuel-gauge light, it would also be useful for the car to possess a database of its own capacities and limits. Such self-monitoring (C2) would include an integrated image of itself—including its current location and fuel consumption, for example—as well as its internal databases (such as "knowing" that it possesses a GPS map that can locate gas stations). A self-monitoring machine would keep a list of its subprograms, compute estimates of their probabilities of succeeding at various tasks, and constantly update them (for example, noticing when a part fails).

Most present-day machine-learning systems are devoid of any self-monitoring; they compute (C0) without representing the extent and limits of their knowledge or the fact that others may have a different viewpoint than their own. There are a few exceptions: Bayesian networks (*54*) or programs (*73*) compute with probability distributions and therefore keep track of how likely they are to be correct. Even when the primary computation is performed by a classical CNN, and is therefore opaque to introspection, it is possible to train a second, hierarchically higher neural network to predict the first one's performance (*47*). This approach, in which a system redescribes itself, has been claimed to lead to "the emergence of internal models that are metacognitive in nature and... make it possible for an agent to develop a (limited, implicit, practical) understanding of itself" (*48*). Pathnet (*72*) uses a related architecture to track which internal configurations are most successful at a given task and use this knowledge to guide subsequent processing. Robots have also been programed to monitor their learning progress and use it to orient resources toward the problems that maximize information gain, thus implementing a form of curiosity (*74*).

An important element of C2 that has received relatively little attention is reality monitoring.

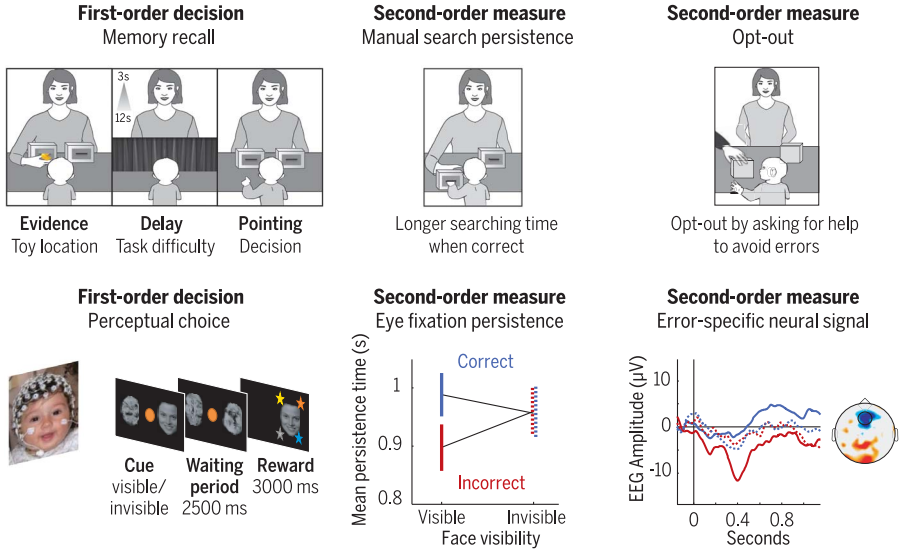


Fig. 3. Self-monitoring: Consciousness in the second sense (C2). Self-monitoring (also called "meta-cognition"), the capacity to reflect on one's own mental state, is available early during infancy. (**Top**) One-and-a-half-year-old infants, after deciding to point to the location of a hidden toy, exhibit two types of evidence for self-monitoring of their decision. (i) They persist longer in searching for the hidden object within the selected box when their initial choice was correct than when it was incorrect. (ii) When given the opportunity to ask for help, they use this option selectively to reduce the probability of making an error. (**Bottom**) One-year-old infants were presented with either a meaningless pattern or a face that was either visible or invisible (depending on its duration) and then decided to gaze left or right in anticipation of face reappearance. As for manual search, post-decision persistence in waiting at the same gaze location increased for correct compared with incorrect initial decisions. Moreover, EEG signals revealed the presence of the error-related negativity over fronto-central electrodes when infants make an incorrect choice. These markers of metacognition were elicited by visible but not by invisible stimuli, as also shown in adults (*61*).

Bayesian approaches to AI (*2*, *73*) have recognized the usefulness of learning generative models that can be jointly used for actual perception (present), prospective planning (future), and retrospective analysis (past). In humans, the same sensory areas are involved in both perception and imagination. As such, some mechanisms are needed to tell apart self-generated versus externally triggered activity. A powerful method for training generative models, called adversarial learning (*75*), involves having a secondary network "compete" against a generative network so as to critically evaluate the authenticity of self-generated representations. When such reality monitoring (C2) is coupled with C1 mechanisms, the resulting machine may more closely mimic human consciousness in terms of affording global access to perceptual representations while having an immediate sense that their content is a genuine reflection of the current state of the world.

Concluding remarks

Our stance is based on a simple hypothesis: What we call "consciousness" results from specific types of information-processing computations, physically realized by the hardware of the brain. It differs from other theories in being resolutely computational; we surmise that mere information-theoretic quantities (*76*) do not suffice to define consciousness unless one also considers the nature and depth of the information being processed.

We contend that a machine endowed with C1 and C2 would behave as though it were conscious; for instance, it would know that it is seeing something, would express confidence in it, would report it to others, could suffer hallucinations when its monitoring mechanisms break down, and may even experience the same perceptual illusions as humans. Still, such a purely functional definition of consciousness may leave some readers unsatisfied. Are we "over-intellectualizing" consciousness, by assuming that some high-level cognitive functions are necessarily tied to consciousness? Are we leaving aside the experiential component ("what it is like" to be conscious)? Does subjective experience escape a computational definition?

Although those philosophical questions lie beyond the scope of the present paper, we close by noting that empirically, in humans the loss of C1 and C2 computations covaries with a loss of subjective experience. For example, in humans, damage to the primary visual cortex may lead to a neurological condition called "blindsight," in which the patients report being blind in the affected visual field. Remarkably, those patients can localize visual stimuli in their blind field but cannot report them (C1), nor can they effectively assess their likelihood of success (C2)—they believe that they are merely "guessing." In this example, at least, subjective experience appears to cohere with possession of C1 and C2. Although centuries of philosophical dualism have led us to consider consciousness as unreducible to physical interactions, the empirical evidence is compatible with the possibility that consciousness arises from nothing more than specific computations.

REFERENCES AND NOTES

1. D. Silver *et al*., *Nature* **529**, 484–489 (2016).
2. B. M. Lake, T. D. Ullman, J. B. Tenenbaum, S. J. Gershman, *Behav. Brain Sci.* **2016**, 1–101 (2016).
3. J. Jaynes, *The Origin of Consciousness in the Breakdown of the Bicameral Mind* (Houghton Mifflin Company, 1976).
4. S. Kouider, E. Dupoux, *Psychol. Sci.* **16**, 617–625 (2005).
5. S. Kouider, S. Dehaene, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **362**, 857–875 (2007).
6. E. Qiao *et al*., *Neuroimage* **49**, 1786–1799 (2010).
7. N. Faivre, L. Mudrik, N. Schwartz, C. Koch, *Psychol. Sci.* **25**, 2006–2016 (2014).
8. S. Kouider, S. Dehaene, *Exp. Psychol.* **56**, 418–433 (2009).
9. S. J. Luck, E. K. Vogel, K. L. Shapiro, *Nature* **383**, 616–618 (1996).
10. S. van Gaal *et al*., *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **369**, 20130212 (2014).
11. H. C. Lau, R. E. Passingham, *J. Neurosci.* **27**, 5805–5811 (2007).
12. S. van Gaal, V. A. Lamme, J. J. Fahrenfort, K. R. Ridderinkhof, *J. Cogn. Neurosci.* **23**, 91–105 (2011).
13. F. P. de Lange, S. van Gaal, V. A. Lamme, S. Dehaene, *PLOS Biol.* **9**, e1001203 (2011).
14. D. Vorberg, U. Mattler, A. Heinecke, T. Schmidt, J. Schwarzbach, *Proc. Natl. Acad. Sci. U.S.A.* **100**, 6275–6280 (2003).
15. S. Dehaene *et al*., *Nature* **395**, 597–600 (1998).
16. A. Vlassova, C. Donkin, J. Pearson, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 16214–16218 (2014).
17. M. Pessiglione *et al*., *Neuron* **59**, 561–567 (2008).
18. M. Pessiglione *et al*., *Science* **316**, 904–906 (2007).
19. L. Polansky, W. Kilian, G. Wittemyer, *Proc. Biol. Sci.* **282**, 20143042 (2015).
20. B. Baars, *A Cognitive Theory of Consciousness* (Cambridge Univ. Press, 1988).
21. S. Dehaene, M. Kerszberg, J. P. Changeux, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14529–14534 (1998).
22. D. Dennett, *Cognition* **79**, 221–237 (2001).
23. S. Dehaene, L. Naccache, *Cognition* **79**, 1–37 (2001).
24. W. James, *The Principles of Psychology* (Holt, 1890).
25. L. Naccache, E. Blandin, S. Dehaene, *Psychol. Sci.* **13**, 416–424 (2002).
26. R. W. Kentridge, C. A. Heywood, L. Weiskrantz, *Proc. Biol. Sci.* **266**, 1805–1811 (1999).
27. C. L. Asplund, D. Fougnie, S. Zughni, J. W. Martin, R. Marois, *Psychol. Sci.* **25**, 824–831 (2014).
28. E. Vul, D. Hanus, N. Kanwisher, *J. Exp. Psychol. Gen.* **138**, 546–560 (2009).
29. R. Moreno-Bote, D. C. Knill, A. Pouget, *Proc. Natl. Acad. Sci. U.S.A.* **108**, 12491–12496 (2011).
30. E. Vul, M. Nieuwenstein, N. Kanwisher, *Psychol. Sci.* **19**, 55–61 (2008).
31. T. I. Panagiotaropoulos, G. Deco, V. Kapoor, N. K. Logothetis, *Neuron* **74**, 924–935 (2012).
32. N. K. Logothetis, *Philos. Trans. R. Soc. Lond. B Biol. Sci.* **353**, 1801–1818 (1998).
33. C. Sergent, S. Baillet, S. Dehaene, *Nat. Neurosci.* **8**, 1391–1400 (2005).
34. S. Marti, M. Sigman, S. Dehaene, *Neuroimage* **59**, 2883–2898 (2012).
35. S. Marti, J.-R. King, S. Dehaene, *Neuron* **88**, 1297–1307 (2015).
36. M. Nieuwenstein, E. Van der Burg, J. Theeuwes, B. Wyble, M. Potter, *J. Vis.* **9**, 1–14 (2009).
37. G. Kreiman, I. Fried, C. Koch, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 8378–8383 (2002).
38. R. Q. Quiroga, R. Mukamel, E. A. Isham, R. Malach, I. Fried, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 3599–3604 (2008).
39. J.-R. King, N. Pescetelli, S. Dehaene, *Neuron* **92**, 1122–1134 (2016).
40. D. Trübutschek *et al*., *eLife* **6**, e23871 (2017).
41. S. Kouider *et al*., *Science* **340**, 376–380 (2013).
42. R. T. Maloney, J. Jayakumar, E. V. Levichkina, I. N. Pigarev, T. R. Vidyasagar, *Exp. Brain Res.* **228**, 365–376 (2013).
43. K. Watanabe, S. Funahashi, *Nat. Neurosci.* **17**, 601–611 (2014).
44. G. N. Elston, *Cereb. Cortex* **13**, 1124–1138 (2003).
45. F.-X. Neubert, R. B. Mars, A. G. Thomas, J. Sallet, M. F. S. Rushworth, *Neuron* **81**, 700–713 (2014).
46. L. Wang, L. Uhrig, B. Jarraya, S. Dehaene, *Curr. Biol.* **25**, 1966–1974 (2015).
47. A. Cleeremans, B. Timmermans, A. Pasquali, *Neural Netw.* **20**, 1032–1039 (2007).
48. A. Cleeremans, *Cogn. Sci.* **38**, 1286–1315 (2014).
49. J. Dunlosky, J. Metcalfe, *Metacognition* (Sage Publications, 2008).
50. A. Clark, A. Karmiloff-Smith, *Mind Lang.* **8**, 487–519 (1993).
51. F. Meyniel, D. Schlunegger, S. Dehaene, *PLOS Comput. Biol.* **11**, e1004305 (2015).
52. F. Meyniel, S. Dehaene, *Proc. Natl. Acad. Sci. U.S.A.* **114**, E3859–E3868 (2017).
53. J. D. Smith, *Trends Cogn. Sci.* **13**, 389–396 (2009).
54. W. J. Ma, J. M. Beck, P. E. Latham, A. Pouget, *Nat. Neurosci.* **9**, 1432–1438 (2006).
55. S. M. Fleming, R. S. Weil, Z. Nagy, R. J. Dolan, G. Rees, *Science* **329**, 1541–1543 (2010).
56. K. Miyamoto *et al*., *Science* **355**, 188–193 (2017).
57. A. Kepecs, N. Uchida, H. A. Zariwala, Z. F. Mainen, *Nature* **455**, 227–231 (2008).
58. E. Rounis, B. Maniscalco, J. C. Rothwell, R. E. Passingham, H. Lau, *Cogn. Neurosci.* **1**, 165–175 (2010).
59. A. Resulaj, R. Kiani, D. M. Wolpert, M. N. Shadlen, *Nature* **461**, 263–266 (2009).
60. L. Charles, J.-R. King, S. Dehaene, *J. Neurosci.* **34**, 1158–1170 (2014).
61. L. Charles, F. Van Opstal, S. Marti, S. Dehaene, *Neuroimage* **73**, 80–94 (2013).
62. C. D. Frith, *The Cognitive Neuropsychology of Schizophrenia* (Psychology Press, 1992).
63. J. S. Simons, J. R. Garrison, M. K. Johnson, *Trends Cogn. Sci.* **21**, 462–473 (2017).
64. D. Mendoza-Halliday, J. C. Martinez-Trujillo, *Nat. Commun.* 10.1038/ncomms15471 (2017).
65. L. Goupil, S. Kouider, *Curr. Biol.* **26**, 3038–3045 (2016).
66. L. Goupil, M. Romand-Monnier, S. Kouider, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 3492–3496 (2016).
67. G. D. Logan, M. J. Crump, *Science* **330**, 683–686 (2010).
68. S. Nieuwenhuis, K. R. Ridderinkhof, J. Blom, G. P. Band, A. Kok, *Psychophysiology* **38**, 752–760 (2001).
69. B. Bahrami *et al*., *Science* **329**, 1081–1085 (2010).
70. R. S. Siegler, *J. Exp. Psychol. Gen.* **117**, 258–275 (1988).
71. V. de Gardelle, P. Mamassian, *Psychol. Sci.* **25**, 1286–1288 (2014).
72. C. Fernando, D. Banarse, C. Blundell, Y. Zwols, D. Ha, A. A. Rusu, A. Pritzel, D. Wierstra, arXiv:170108734 [cs.NE] (2017).
73. J. B. Tenenbaum, C. Kemp, T. L. Griffiths, N. D. Goodman, *Science* **331**, 1279–1285 (2011).
74. J. Gottlieb, P.-Y. Oudeyer, M. Lopes, A. Baranes, *Trends Cogn. Sci.* **17**, 585–593 (2013).
75. I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative Adversarial Networks. arXiv:14062661 [stat.ML] (2014).
76. G. Tononi, M. Boly, M. Massimini, C. Koch, *Nat. Rev. Neurosci.* **17**, 450–461 (2016).
77. S. Kouider, E. Eger, R. Dolan, R. N. Henson, *Cereb. Cortex* **19**, 13–23 (2009).
78. S. Dehaene *et al*., *Nat. Neurosci.* **4**, 752–758 (2001).
79. P. Vuilleumier *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **98**, 3495–3500 (2001).
80. A. Kiesel, W. Kunde, C. Pohl, M. P. Berner, J. Hoffmann, *J. Exp. Psychol. Learn. Mem. Cogn.* **35**, 292–298 (2009).
81. M. Aly, A. P. Yonelinas, *PLOS ONE* **7**, e30231 (2012).
82. I. M. Harlow, A. P. Yonelinas, *Memory* **24**, 114–127 (2016).
83. H. L. Pincham, H. Bowman, D. Szucs, *Cortex* **81**, 35–49 (2016).
84. C. Sergent, S. Dehaene, *Psychol. Sci.* **15**, 720–728 (2004).
85. A. Del Cul, S. Baillet, S. Dehaene, *PLOS Biol.* **5**, e260 (2007).
86. R. Marois, D. J. Yi, M. M. Chun, *Neuron* **41**, 465–472 (2004).
87. C. Moutard, S. Dehaene, R. Malach, *Neuron* **88**, 194–206 (2015).
88. H. G. Rey, I. Fried, R. Quian Quiroga, *Curr. Biol.* **24**, 299–304 (2014).
89. A. Schurger, I. Sarigiannidis, L. Naccache, J. D. Sitt, S. Dehaene, *Proc. Natl. Acad. Sci. U.S.A.* **112**, E2083–E2092 (2015).
90. C. Sergent *et al*., *Curr. Biol.* **23**, 150–155 (2013).
91. J. Sackur, S. Dehaene, *Cognition* **111**, 187–211 (2009).
92. R. Marois, J. Ivanoff, *Trends Cogn. Sci.* **9**, 296–305 (2005).
93. P. Barttfeld *et al*., *Proc. Natl. Acad. Sci. U.S.A.* **112**, 887–892 (2015).
94. W. J. Gehring, B. Goss, M. G. H. Coles, D. E. Meyer, E. Donchin, *Psychol. Sci.* **4**, 385–390 (1993).

ACKNOWLEDGMENTS

This work was supported by funding from Institut National de la Santé et de la Recherche Médicale, CEA, Canadian Institute for Advanced Research, Spoelberch Foundation, Collège de France, the European Research Council (ERC project NeuroConsc to S.D. and ERC project METAWARE to S.K), the French Agence National de la Recherche (grants ANR-10-LABX-0087 and ANR-10-IDEX-0001-02), the U.S. National Institute of Health (National Institute of Neurological Disorders and Stroke grant R01NS088628 to H.L.), and the U.S. Air Force Office of Scientific Research (grant FA9550-15-1-0110 to H.L.)

10.1126/science.aan8871

What is consciousness, and could machines have it?

Stanislas Dehaene, Hakwan Lau and Sid Kouider

Science **358** (6362), 486-492.
DOI: 10.1126/science.aan8871