the Pontifical Academy of Sciences

What is Consciousness, and Could Machines Have It?

Stanislas Dehaene

Although consciousness is frequently considered as the pinnacle of the brain, and something that is impossible to confer to machines, I would like to argue otherwise. In this report, I argue that, in fact, much is already known about how brains generate consciousness, and how those findings could be used in artificial intelligence. In the past twenty years, cognitive neuroscience has made great advances in understanding the "signatures of consciousness" – brain activity markers that occur only when the subject is aware of a certain content. Those findings support a theory of consciousness as a sharing device, a "global neuronal workspace" that allows us to share our knowledge, internally, between all of the specialized "processors" in our brain, and externally, with other people. On this basis, I make several tentative suggestions as to which functionalities should be added to present-day machines before they might be considered conscious.

The contemporary strategy to study consciousness

In the past 20 years, the problem of consciousness has ceased to appear insurmountable. Neuroscience has started to identify the objective brain mechanisms underlying subjective processes – what I call the "signatures" of conscious processing. Those discoveries have been reviewed in detail elsewhere (Dehaene & Changeux, 2011; Dehaene, 2014). Briefly, the first advance came with the advent of explicit theories of computation (Hilbert, Gödel, Turing, Von Neumann) and information representation (Shannon). Following those breakthroughs, consciousness could then be seen as a computational property associated with a certain level of information processing.

At present, three computational levels may be distinguished. At the lowest level, which we may call level 0, unconscious algorithms process symbols blindly and, obviously, without any awareness. For instance, our visual system blindly and unconsciously processes the following image (due to Adelson).

It lets us see a strictly normal checkerboard, although this is illusory – as you may check by masking the figure, the two knights and their squares seem to be black and white, but they are actually exactly the same shade of grey. What is happening? Our visual system detects the presence of a dark zone in the image, which it interprets as a shadow, and it subtracts it from the image to let us see the "true" shade of grey of the pieces, thus making one knight look brighter than the other. Arguably, any machine whose aim would be to extract the genuine appearance of objects while getting rid of shadows and other defects of the image would have to go through the same inference process. In this sense, many of the brain's unconscious computations are rational computations. Paradoxically, any machine that strives towards objectivity would be submitted to similar human-like illusions.

Above the unconscious processing level, two higher levels of information processing may be defined, corresponding to what others have termed primary and secondary consciousness (Edelman, 1989).

- Level 1 is conscious access. At any given moment, although our brain is bombarded with stimuli and has a vast repertoire of possible sensory or memory states, only a single piece of information, selected for its relevance, is consciously accessed, amplified, and becomes the focus of additional processing. This selective attribution of higher-level computing resources is what we experience as "conscious access".

- Level 2 is conscious self-representation. At this level, the cognitive system entertains one or several representations of its own knowledge, for instance it may know what it is currently focusing on, that it made an error, etc. Thus, the system not only commits its resources to a specific piece of information (level 1), but also "knows that it knows" (level 2). The assumption is that this self-knowledge is represented in the same format as the knowledge of other people (also known as a "theory of mind"), thus allowing this information to be shared with others and to be used in social decision making (Bahrami et al., 2010; Frith, 2007).

Baars (1989) was one of the first cognitive scientists to realize that, given those simple definitions, it is quite possible to study consciousness experimentally. The experimental strategy proceeds in several steps (Dehaene, 2014):

1. Identify a minimal experimental paradigm (e.g. a visual illusion) that allows to contrast visible and invisible stimuli. My laboratory has used masking, whereby a flashed image can be made either subliminal or conscious

(Kouider & Dehaene, 2007). Others have used binocular rivalry, whereby the competition between two images is used to render one of them conscious while the other is not (Logothetis, Leopold, & Sheinberg, 1996). Many other minimal contrasts are available, for instance sleep versus wakefulness; wakefulness versus anesthesia; vegetative-state versus minimally conscious patients, etc. (Baars, 1989).

2. Carefully quantify the subject's introspection, i.e. what he or she "knows that it knows". Introspection defines the very phenomenon that we want to study (conscious subjective perception) and must therefore be recorded alongside other objective measures of behavior and brain activity. The ideal situation consists in presenting a fixed stimulus closed to the threshold for awareness, and to sort the trials according to subjective reports, such that the very same stimulus is sometimes perceived consciously and sometimes remains unconscious.

3. As a consequence, focus on a particular and restricted sense of consciousness: the capacity to report a piece of information, to oneself or to others. Scientists have learned that this sense of consciousness, called *reportability*, is well-defined and differs from other concepts such as attention, vigilance, or self-consciousness.

4. Apply the panoply of modern neuro-imaging and neuroscience tools to compare the behaviors and brain activity patterns evoked by reportable and unreportable stimuli, thus uncovering the signatures of consciousness.

Current signatures of consciousness in the human brain

Experiments that have implemented this strategy have discovered that, although subliminal stimuli can induce considerable activity in many if not all circuits of the human brain, conscious perception is associated with a set of specific signatures:

- **Amplification and access to prefrontal cortex.** Compared to a subliminal image, a conscious image is amplified and gains access to higher levels of representation, particularly in prefrontal and parietal cortices.

- **Late global ignition and meta-stability**. Tracking the propagation of conscious and unconscious images shows that unconscious activity can be strong in early visual cortex, yet die out in a few hundreds of milliseconds within higher cortical areas. A conscious image, on the contrary, is amplified in a non-linear manner, an event called "global ignition". By about 300 milliseconds, brain activity becomes more stable when the stimulus is conscious than when it is not (Schurger, Sarigiannidis, Naccache, Sitt, & Dehaene, 2015).

- **Brain-scale diffusion of information.** Conscious ignition is accompanied by increased in bidirectional exchanges of information in the human brain. During a conscious episode, the cortex "talks to itself" at greater distances, and this is manifested by correlations of brain signals, particularly in the beta band (13-30 Hz) and theta band (3-8 Hz).

- **Global spontaneous activity.** Even in the absence of stimuli, the brain spontaneously generates its own patterns of distributed activity, which are constantly changing (Barttfeld et al., 2015). This resting state activity can partially predict the content of consciousness, for instance whether the subject currently experiences mental images or "mind wandering".

- **Late all-or-none firing of "concept cells".** Single-cell correlates of conscious ignition have been identified in human and non-human primates. Neurons in prefrontal and anterior temporal cortex fire to a specific concept (e.g. the Empire State building) and do so *only* when the corresponding word or image is presented consciously. Their late activity acts as a signature of conscious perception (Quiroga, Mukamel, Isham, Malach, & Fried, 2008).

Those findings are compatible with the Global Neuronal Workspace (GNW) hypothesis, a simple theory of consciousness (Baars, 1989; Dehaene & Changeux, 2011; Dehaene, 2014). Briefly, the hypothesis is that, while specialized subsystems of the brain ("modules") process information unconsciously, what we subjectively experience as consciousness is the global availability of information, which is made possible by a non-modular "global workspace". Consciousness is a computational device that evolved to break the modular organization of the brain. During conscious perception, a distributed parieto-frontal circuit, forming the global neuronal workspace, ignites to selectively amplify a relevant piece of information. Thanks to its long-distance connectivity, supported by giant neurons with long axons in layers 2-3, it stabilizes a selected piece of information and broadcasts it in a brain-wide manner to all other modules. The global workspace thus maintains information in an active state for as long as it is needed (meta-stability).

The GNW hypothesis addresses the classical question of the function of consciousness. Is consciousness a mere epiphenomenon, i.e. a useless side-effect of brain activity, similar to the whistle of the train? Theory and experiments suggest otherwise: consciousness appears to be required for specific operations. Thanks to the global workspace, we can reflect upon the information: subliminal information is evanescent, but conscious information is stabilized and available for long-term thinking. Consciousness is also helpful in order to discretize

the incoming flux of information and reduce it to a few samples that can be reported or stored: while unconscious processes compute with an entire probability distribution, consciousness samples from it. Consciousness is also involved in routing information to other processing stages, thus allowing us to perform arbitrary chains of operations (for instance, computing 23x47; Sackur & Dehaene, 2009). Finally, consciousness plays a key role in monitoring our behavior and diagnosing our errors. We have found that a key component of the brain's error monitoring system, the "error negativity" that arises whenever we press the wrong button in a simple response-time task, only occurs on trials where subjects report seeing the stimulus (Charles, Van Opstal, Marti, & Dehaene, 2013). Only visible stimuli allow us to detect the occasional discrepancy between what we intended and what we did.

What machines are missing

In summary, cognitive neuroscientists are beginning to understand that the computations that we experience as "conscious processing" are useful aspects of brain function that are therefore likely to be equally useful to artificial-intelligence devices. Here, I list four computational features that conscious brains possess and that machines currently miss. My suggestion is that if those functions were implemented, the resulting machine would be likely to be considered conscious, or at least much closer to conscious than most machines currently are.

1. **A workspace for global information sharing.** In current computers and cellphones, computations are performed by special-purpose programs known as "apps". Each app possesses its own memory space and its specific knowledge base, carefully protected from others. Apps do not share their knowledge: it is frequent for one app to "know" a piece of information while others ignore it. In the human brain, this is characteristic of unconscious processing. According to the GNW hypothesis, consciousness evolved to break this modularity. The GNW can extract relevant information from virtually any brain module, and make it available to the entire organism. Machines may benefit from a similar architecture for flexible information sharing, capable of broadcasting to the entire system a potentially relevant piece of information. "Blackboard" architectures of this type were proposed in the 1970s. It would be interesting to pursue this idea in the context of present-day machine-learning algorithms, which are able to make the best use of the broadcasted information.

2. **A repertoire of self-knowledge.** To determine where the important information lies and where to route it, I believe that brains and machines alike must be endowed with a repertoire of self-knowledge. By this, I do not mean a bodily sense of self, as might be available for instance to a robot that would know the location of its limbs (in the human brain, the construction of this body map is, in fact, unconscious). What I have in mind is an internal representation of the machine's own abilities: a database that contains a list of its apps, the kind of knowledge they possess, what goals they can fulfill, how fast they can operate, how likely they are to be correct, etc. Even young children, when learning arithmetic, compile such a repertoire of the different strategies at their disposal (Siegler & Jenkins, 1989). In a machine endowed with learning algorithms, self-knowledge should be constantly updated, leading to a concrete implementation of the Socratic "know thyself".

3. **Confidence and "knowing that you don't know".** A conscious machine should know when it is wrong or when it is uncertain about something. In the human brain, this corresponds to meta-cognitive knowledge (cognition about cognition) which has been linked to prefrontal cortex. Even preverbal infants know that they don't know, as revealed by the fact that they turn to their mother for help whenever appropriate {Kouider}. There are several ways in which a computer could be equipped with a similar functionality. First, it could be endowed with statistical programs that do not just give an answer, but also compute the probability that this answer is correct (according to Bayes' law or some approximation of it). Second, a computer could be endowed with an error-detection system, similar to the brain's error-negativity, which constantly compares ongoing activity with prior expectations and spontaneously reacts if the current behavior is likely to be wrong. Third, this error-detection device could be coupled to a corrective device, such that the system constantly looks for alternative ways to get the correct answer.

4. **Theory of mind and relevance.** One aspect of consciousness, which may be unique to humans, is the ability to represent self-knowledge in the same format as knowledge of others. Every human being holds distinct representations of what he knows; what others know; what he knows that others know; what he knows that others don't know; what others know that he doesn't know; etc. This faculty, called theory of mind, is what allows us to model other minds and to use this knowledge in order to maximize the usefulness of information that we can provide them (relevance, as defined by Sperber & Wilson, 1988). Current machines often lack such relevance. A machine that could simulate its user's mind would undoubtedly provide more relevant information. It would remember what it previously said, infer what its user knows, and avoid presenting trivial, useless, or otherwise contradictory information. Algorithms that handle such recursive representations of other minds are currently being developed (Baker, Saxe, & Tenenbaum, 2009; Daunizeau et al., 2010).

The above list is probably not exhaustive. However, I contend that conferring it to computers would arguably go a long way towards closing the consciousness gap. According to the present stance, consciousness is not an essence, but solely a functional property that can be progressively approximated. Humans seem to be quite generous in attributing consciousness to others, including animals, plants, and even inanimate objects such as clouds or storms. The steps that I outlined here should bring us closer to attributing consciousness to machines.

References

Baars, B. (1989). *A cognitive theory of consciousness*. Cambridge, Mass.: Cambridge University Press.

Bahrami, B., Olsen, K., Latham, P.E., Roepstorff, A., Rees, G., & Frith, C.D. (2010). Optimally interacting minds. *Science*, *329*(5995), 1081-5. https://doi.org/10.1126/science.1185718

Baker, C.L., Saxe, R., & Tenenbaum, J.B. (2009). Action understanding as inverse planning. *Cognition*, *113*(3), 329-49. https://doi.org/10.1016/j.cognition.2009.07.005

Barttfeld, P., Uhrig, L., Sitt, J.D., Sigman, M., Jarraya, B., & Dehaene, S. (2015). Signature of consciousness in the dynamics of resting-state brain activity. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 887-892. https://doi.org/10.1073/pnas.1418031112

Charles, L., Van Opstal, F., Marti, S., & Dehaene, S. (2013). Distinct brain mechanisms for conscious versus subliminal error detection. *Neuroimage*, *73*, 80-94. https://doi.org/10.1016/j.neuroimage.2013.01.054

Daunizeau, J., den Ouden, H.E., Pessiglione, M., Kiebel, S.J., Stephan, K.E., & Friston, K.J. (2010). Observing the observer (I): meta-bayesian models of learning and decision-making. *PLoS One*, *5*(12), e15554. https://doi.org/10.1371/journal.pone.0015554

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts* (Reprint edition). Penguin Books.

Dehaene, S., & Changeux, J.P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*(2), 200–27. https://doi.org/10.1016/j.neuron.2011.03.018

Edelman, G. (1989). *The remembered present.* Basic Books: New York.

Frith, C. (2007). *Making up the mind. How the brain creates our mental world*. London: Blackwell.

Kouider, S., & Dehaene, S. (2007). Levels of processing during non-conscious perception: a critical review of visual masking. *Philos Trans R Soc Lond B Biol Sci*, *362*(1481), 857-75.

Logothetis, N.K., Leopold, D.A., & Sheinberg, D.L. (1996). What is rivalling during binocular rivalry? *Nature*, *380*(6575), 621-4.

Quiroga, R.Q., Mukamel, R., Isham, E.A., Malach, R., & Fried, I. (2008). Human single-neuron responses at the threshold of conscious recognition. *Proc Natl Acad Sci U S A*, *105*(9), 3599-604.

Sackur, J., & Dehaene, S. (2009). The cognitive architecture for chaining of two mental operations. *Cognition*, *111*(2), 187-211.

Schurger, A., Sarigiannidis, I., Naccache, L., Sitt, J.D., & Dehaene, S. (2015). Cortical activity is more stable when sensory stimuli are consciously perceived. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(16), E2083-2092. https://doi.org/10.1073/pnas.1418730112

Siegler, R.S., & Jenkins, E.A. (1989). *How children discover new strategies.* Hillsdale N.J.: Lawrence Erlbaum Associates.

Sperber, D., & Wilson, D. (1988). Précis of relevance#: Communication and cognition.*Behavioral and Brain Sciences*, *10*, 697-789.