doi:10.1093/brain/awab046 BRAIN 2021: 144; 1281–1283 | 1281

BRAIN
ESSAY

Understanding consciousness

The word consciousness has different meanings to different people. As a result, the topic is difficult to write about. There are three major perspectives on consciousness, at least as I see them: the spiritual, hard-problem, and self-model perspectives. I argue here that they represent a progression in our understanding. The religious perspective is the original approach, intuitively compelling to many; the hard-problem view represents a conceptual advance introduced towards the end of the 20th century; and the self-model framework, which has gained ground more recently, is entirely mechanistic and, in my view, the most scientifically valid.

The belief in a non-physical spirit is extremely old. Burials with grave goods are thought to be evidence of a belief in a spirit that survives (and is therefore of a different material than) the physical body, and such burials may date back at least a hundred thousand years. Although the belief has an origin in prehistory, perhaps the most famous philosophical formulation of it came from Descartes, who argued that humans are composed of a physical substance (*res extensa*) and an ethereal, spiritual or mental substance (*res cogitans*). This formulation has come to be called Cartesian dualism.

Many modern scientific and philosophical theories of consciousness are a disguised form of dualism, positing that something—some trigger or mechanism in the physical brain—gives rise to a non-physical feeling. The physical mechanism—the neural correlates of consciousness—can be studied, but the non-physical adjunct to it—the feeling itself—is outside the bounds of science. In this form, dualism has influenced modern views of consciousness, such as the hard-problem view discussed next.

William James, sometimes called the founder of modern psychology, coined the term 'stream of consciousness' in the late 1800s.[1] To him, the content of consciousness, constantly changing, constantly flowing, included thoughts, sensory impressions, emotions, decisions, memories, the knowledge that you are a person distinct from the rest of the world, and everything else swirling through your moment-by-moment experience.

If consciousness is a stream of mental content, then perhaps understanding consciousness is a matter of understanding how the brain computes all of that content. As computer technology emerged in the middle part of the 20th century, that mechanistic view of consciousness, as a collection of computed content, began to seem plausible. In 1950, Turing argued that computers could eventually be programmed to think like people.[2]

In the 1970s, however, a philosophical perspective began to emerge, partly as a reaction to the computer analogy. The new perspective was famously put by the philosopher Nagel, in his essay, 'What is it like to be a Bat?'.[3] It was also summarized later by the philosopher Chalmers, who popularized the phrase, 'the hard problem'.[4] In the new philosophical perspective, consciousness is not about the specific content, but rather the subjective experience associated with some of that content.

What is that subjective experience? Why does only some internal content come with a feeling? Psychologists have known for a century that most of the processing in the brain is hidden, occurring without any subjective feeling attached to it. So how does the essence of experience become attached to some items in the brain, and not others? In the hard-problem view, because subjective experience, the 'what it feels like' component, is fundamentally private, because it is not an object with physical properties like mass and hardness, it is also fundamentally outside the realm of objective science. It is undissectable. It, itself, will never be explained. Hence the term 'hard problem', meaning, euphemistically, the impossible-to-solve problem.

The hard-problem perspective categorically separates the non-physical feeling of consciousness from the physical mechanisms in the brain and the information content that is, sometimes, the subject of that feeling. If consciousness is not about having complex information in your head, but rather about having a subjective experience, then a bat could have a subjective experience of its

Received January 19, 2021. Accepted January 29, 2021. Advance access publication March 29, 2021

Image by Andre Mouton, from Pixabay.

world (as Nagel suggested). Or a mouse. Or a bee. Or maybe even a microbe.

A major way in which the hard-problem perspective has impacted neuroscience is through the study of the neural correlates of consciousness. In that approach, scientists study the events in the brain that correlate with the human report of consciousness. If we do so, we have done all we can, scientifically, while the emergent conscious experience itself is not scientifically touchable.

At least one theory of consciousness, the global workspace theory, is consistent with this neural-correlates approach. In the theory, when information in the brain is boosted in signal strength by the mechanisms of attention, it reaches a central network, the global workspace.[5] Once it has entered the global workspace, a subjective feeling emerges from that information. The person becomes conscious of it. One can study information, attention, and the global workspace, but as to the subjective experience that emerges from it . . . it simply appears.

A second theory of consciousness, the higher-order thought theory, posits that when low-level sensory information is incorporated into higher-order thoughts, then a subjective feeling of consciousness arises.[6] We can scientifically understand how information moves from low-level to higher-order, but in the end, a fundamentally irreducible phenomenon occurs—the subjective feeling.

A third currently popular theory, the integrated information theory, posits that when a system contains information, and when that information is 'integrated' together according to a mathematical metric, then the feeling of consciousness arises.[7] We can scientifically describe the conditions under which the conscious feeling occurs—but the feeling itself remains irreducible.

All of these theories, and many others, stem from an often-unspoken philosophical perspective, a dualist perspective. In that perspective, the brain is a physical engine that can be understood scientifically; the engine generates a fundamentally non-physical product, a subjective feeling or experience; and that product is beyond further understanding or mechanistic deconstruction.

The final perspective that I'll describe, the self-model perspective, is, in my view, the next natural step. In the hard-problem perspective, merely processing information—such as processing the colour of an apple—does not explain conscious experience. Something extra must be present, a non-physical essence, a feeling that science cannot further deconstruct. In the self-model view, once again, merely processing information about the colour of an apple is not enough for conscious experience of the apple. Something else must be present. However, the extra ingredient is mechanistically understandable and plays a useful role in cognition.

To understand how the self-model approach works, first consider a small, but crucial, piece of logic. Everything that you think is true about yourself—everything, no matter how certain you are of it—stems from information in the brain, or you wouldn't be able to think the thought or articulate the claim. The brain is a model builder. It builds models, or bundles of information, descriptive of things in the world. The visual system builds visual models, rich sets of information that represent objects. The body schema is a set of information that represents the physical structure and state of the body. Our general beliefs about the world are models at a more cognitive level.

Where does that leave us with respect to consciousness?

Most people are absolutely certain they have an ineffable, subjective feeling that accompanies their thinking and their perception. The standard argument could be put this way: 'I know I have it, because I'm experiencing it right now. I can feel it'. But this argument is a classical tautology, equivalent to saying, 'I know I have it because I have it; I know it's true because it's true. I know I have a feeling, because I feel the feeling'.

What has happened here? Normally, if a person is absolutely certain of something—whether plausible or incoherent—we understand how to interpret the situation. The person's brain has constructed a set of information, on the basis of which the belief and the certainty occurs. In the case of the subjective feeling, the brain is stuck in a logic loop. Cognition has gained access to an information set; the information set is part of a self-model; it is information that describes some aspect of the self. On the basis of that information, cognition arrives at the belief and the certainty that a non-physical feeling is present. But that self-model is unlikely to be a literally accurate representation. We think we have something non-physical and intangible inside us, because, whatever it is that we actually have, whatever physical process is the subject of that self-model, the model depicts it in an incomplete manner.

This perspective is sometimes called illusionism, because conscious experience is said to be an illusion.[8] I have argued that the term illusionism is not right. It gives the impression that nobody is home and nothing exists in our heads. But that isn't correct. Our belief that we have a conscious feeling inside us—along with every other belief and conviction and perception and thought that we have—derives from information in the brain. The scientific question of consciousness becomes: what is that crucial information set, the self-model on which our belief in a hard problem of consciousness depends?

The attention schema theory (AST) is, to my knowledge, the most fully elaborated version of a self-model theory of consciousness.[9] AST focuses on the close relationship between subjective awareness and objective attention. Attention is a mechanism by which some items are given a signal boost in the brain, and are thereby processed in greater depth and gain a greater influence over output systems. Attention moves and changes, capturing different items over time. It can be focused on internal thoughts and memories, just as well as on external stimuli. It is the tool through which the brain concentrates resources on whatever seems most relevant at the moment.

Though attention may seem like merely one of many processes in the brain, it is of special importance. It is, arguably, the key to any kind of complex intelligence, because it allows the limited resources of the brain to process selected items in great depth and arrive at a complex response. It is an entirely mechanistic, physical process, so much so that many artificial intelligence systems have had versions of attention built into them.

Subjective awareness is similar in many ways. When we are aware of something, we feel that we're processing it, we grasp it with the mind, we're able to respond to it—in these respects,

awareness and attention sound similar. Moreover, attention and awareness almost always move together. What your brain is attending to, you're almost always subjectively aware of.

Attention and awareness are so closely associated that it's tempting to think they may be different labels for the same process. However, it turns out they can be separated. Under laboratory circumstances, it is possible to manipulate people to pay attention to a visual stimulus (in the sense of focusing the brain's processing resources on it) without any subjective awareness of the stimulus. If the stimulus is very dim or brief, attention will flicker towards it, while the person will insist that nothing appeared.

What can we make of this strange relationship between awareness and attention, two processes that appear to act the same way most of the time, such that one of them is evidently redundant, but then occasionally they slip and act separately? What does it mean that attention is a physical, objectively measurable process in the brain, whereas awareness is a property that we only 'know' about personally and attest to?

AST makes simple sense of this complex pattern. In the theory, the brain controls attention with the help of a model of attention. That model, or 'attention schema', is a constantly updating set of information that describes the current state of attention and predicts how attention may transition into future states. Without the model, the brain can't steer attention. In a similar manner, the motor system controls the arm with the help of a descriptive and predictive model of the arm, a part of the body schema. A general principle of control systems is that, to be good at controlling something, the system needs a useful model of the thing it controls.

According to AST, when we claim to have subjective awareness of something, the claim stems from the information in that attention schema. For example, when you look at an apple, your knowledge about the apple—it's colour and shape—comes from a sensory model constructed in the visual system. But your belief that there is something else, a subjective experience, a feeling that comes with processing the apple—that belief stems from an attention schema, a model of the process of attention.

Why do we believe that subjective awareness is a non-physical essence, a hard problem? Because we are misled by that model of attention. Being an imperfect model, lacking details, its depiction of attention is of a mysterious, non-physical essence that can seize hold of items and vividly know them. Does that mean that consciousness is an illusion? No, it means that the magic essence we think we have is a useful caricature. It acts as a model that the brain needs to function.

One of the challenges in consciousness research is that being conscious of something does not have many definite, demonstrable benefits. Does consciousness have any purpose? Do people have any ability that, under laboratory tests, depends on the presence of a subjective experience? AST says yes, and the data back it up. One such ability, it turns out, is the control of attention.[10] If you're not aware of a bug flying at the edge of your vision, your attention may flicker to that bug, pulled there involuntarily, but your endogenous control over that attention is crippled. You will be unable to suppress attention, sustain attention, or strategically shift attention with respect to that bug.

Without awareness (without the control model telling the brain what attention is doing), the control of attention collapses. When the control of attention collapses, our behaviour collapses. After all, creating a cognitive plan and executing it requires a controlled, sequential movement of attention from one item to the next. The very essence of intelligent agency depends on the control of attention, and therefore on awareness.

Several implications follow from AST. First, subjective experience is not just a bonus gift, a happy side effect of the brain. It has a specific, adaptive function—it enables us to control our own attention. Second, it must have begun to evolve when animals developed a sophisticated ability to control their own attention, probably a few hundred million years ago. Mammals and birds, possibly reptiles, must share a similar basic mechanism. Third, in the human brain, we can point to networks that are probably responsible for it, networks that are known to be involved in controlling attention. They are exactly the networks that, when damaged, lead to general disruptions in awareness, such as the clinical syndrome of spatial neglect. Finally, for those interested in technology, AST offers a possibility for artificial intelligence. Subjective experience is not a mysterious property that might emerge if you build an artificial neural network big enough. It is a specific, understandable, engineerable trait that serves a useful function.

AST does not necessarily contradict all previous theories of consciousness. Some may have much validity. But they leave the final step unexplained. They leave untouched the question of subjective experience, as though a Cartesian *res cogitans* were rising up from the mechanistic parts of the theory. AST takes that final step. It explains why biological computing machines think they have a magic essence of experience inside them, and why that self-model is crucial to good cognitive function.

Michael S. A. Graziano
Department of Psychology, Princeton University, USA
E-mail: graziano@princeton.edu

Michael Graziano is the author of *Rethinking Consciousness: A Scientific Theory of Subjective Experience*.

doi:10.1093/brain/awab046
Advance access publication March 29, 2021

References

1. James W. *The principles of psychology*. New York, NY: Henry Holt; 1890.
2. Turing AM. Computing machinery and intelligence. *Mind* 1950; LIX:433-460.
3. Nagel T. What is it like to be a bat? *Philos Rev*. 1974;83:435-50.
4. Chalmers D. Facing up to the problem of consciousness. *J Conscious Stud*. 1995;2:200-219.
5. Dehaene S. *Consciousness and the brain*. New York: Viking Press; 2014.
6. Rosenthal D. *Consciousness and mind*. New York: Oxford University Press; 2005.
7. Tononi G, Boly M, Massimini M, Koch C. Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci*. 2016;17:450-461.
8. Dennett DC. *Consciousness explained*. New York: Little-Brown; 1991.
9. Graziano MSA. *Consciousness and the social brain*. New York: Oxford University Press; 2013.
10. Wilterson AI, Kemper CM, Kim N, Webb TW, Reblando AMW, Graziano MSA. Attention control and the attention schema theory of consciousness. *Progr Neurobiol*. 2020;195:101844.