How the light gets out

Consciousness is the ‘hard problem’, the mystery that confounds science and philosophy. Has a new theory cracked it?

by **Michael Graziano**

- Read later or Kindle

Illustration by Michael Marsicano

Scientific talks can get a little dry, so I try to mix it up. I take out my giant hairy orangutan puppet, do some ventriloquism and quickly become entangled in an argument. I’ll be explaining my theory about how the brain — a biological machine — generates consciousness. Kevin, the orangutan, starts heckling me. ‘Yeah, well, I don’t have a brain. But I’m still conscious. What does *that* do to your theory?’

Kevin is the perfect introduction. Intellectually, nobody is fooled: we all know that there’s nothing inside. But everyone in the audience experiences an illusion of sentience emanating from his hairy head. The effect is automatic: being social animals, we project awareness onto the puppet. Indeed, part of the fun of ventriloquism is *experiencing* the illusion while knowing, on an intellectual level, that it isn’t real.

Many thinkers have approached consciousness from a first-person vantage point, the kind of philosophical perspective according to which other people’s minds seem essentially unknowable. And yet, as Kevin shows, we spend a lot of mental energy attributing consciousness to other things. We can’t help it, and the fact that we can't help it ought to tell us something about what consciousness is and what it might be used for. If we evolved to recognise it in others – and to mistakenly attribute it to

puppets, characters in stories, and cartoons on a screen — then, despite appearances, it really can't be sealed up within the privacy of our own heads.

Lately, the problem of consciousness has begun to catch on in neuroscience. How does a brain generate consciousness? In the computer age, it is not hard to imagine how a computing machine might construct, store and spit out the information that 'I am alive, I am a person, I have memories, the wind is cold, the grass is green,' and so on. But how does a brain become *aware* of those propositions? The philosopher David Chalmers has claimed that the first question, how a brain computes information about itself and the surrounding world, is the 'easy' problem of consciousness. The second question, how a brain becomes aware of all that computed stuff, is the 'hard' problem.

I believe that the easy and the hard problems have gotten switched around. The sheer scale and complexity of the brain's vast computations makes the easy problem monumentally hard to figure out. How the brain attributes the property of awareness to itself is, by contrast, much easier. If nothing else, it would appear to be a more limited set of computations. In my laboratory at Princeton University, we are working on a specific theory of awareness and its basis in the brain. Our theory explains both the apparent awareness that we can attribute to Kevin *and* the direct, first-person perspective that we have on our own experience. And the easiest way to introduce it is to travel about half a billion years back in time.

In a period of rapid evolutionary expansion called the Cambrian Explosion, animal nervous systems acquired the ability to boost the most urgent incoming signals. Too much information comes in from the outside world to process it all equally, and it is useful to select the most salient data for deeper processing. Even insects and crustaceans have a basic version of this ability to focus on certain signals. Over time, though, it came under a more sophisticated kind of control — what is now called *attention*. Attention is a data-handling method, the brain's way of rationing its processing resources. It has been found and studied in a lot of different animals. Mammals and birds both have it, and they diverged from a common ancestor about 350 million years ago, so attention is probably at least that old.

Attention requires control. In the modern study of robotics there is something called control theory, and it teaches us that, if a machine such as a brain is to control something, it helps to have an internal model of that thing. Think of a military general with his model armies arrayed on a map: they provide a simple but useful representation — not always perfectly accurate, but close enough to help formulate strategy. Likewise, to control its own state of attention, the brain needs a constantly updated simulation or model of that state. Like the general's toy armies, the model will be schematic and short on detail. The brain will attribute a property to itself and that property will be a simplified proxy for attention. It won't be precisely accurate, but it will convey useful information. What exactly is that property? When it is paying attention to thing X, we know that the brain usually attributes an *experience* of X to itself — the property of being *conscious*, or *aware*, of something. Why? Because that attribution helps to keep track of the ever-changing focus of attention.

The most basic, measurable, quantifiable truth about consciousness is simply this: we humans can say that we have it

I call this the 'attention schema theory'. It has a very simple idea at its heart: that consciousness is a schematic model of one's state of attention. Early in evolution, perhaps hundreds of millions of years ago, brains evolved a specific set of computations to construct that model. At that point, 'I am aware of X' entered their repertoire of possible computations.

And then what? Just as fins evolved into limbs and then into wings, the capacity for awareness probably changed and took on new functions over time. For example, the attention schema might have allowed the brain to integrate information on a massive new scale. If you are attending to an apple, a decent model of that state would require representations of yourself, the apple, and the complicated process of attention that links the two. An internal model of attention therefore collates data from many separate domains. In so doing, it unlocks enormous potential for integrating

information, for seeing larger patterns, and even for understanding the relationship between oneself and the outside world.

Such a model also helps to simulate the minds of other people. We humans are continually ascribing complex mental states — emotions, ideas, beliefs, action plans — to one another. But it is hard to credit John with a fear of something, or a belief in something, or an intention to do something, unless we can first ascribe an awareness of something to him. Awareness, especially an ability to attribute awareness to others, seems fundamental to any sort of social capability.

It is not clear when awareness became part of the animal kingdom's social toolkit. Perhaps birds, with their well-developed social intelligence, have some ability to attribute awareness to each other. Perhaps the social use of awareness expanded much later, with the evolution of primates about 65 million years ago, or even later, with our own genus *Homo*, a little over two million years ago. Whenever it arose, it clearly plays a major role in the social capability of modern humans. We paint the world with perceived consciousness. Family, friends, pets, spirits, gods and ventriloquist's puppets — all appear before us suffused with sentience.

But what about the inside view, that mysterious light of awareness accessible only to our innermost selves? A friend of mine, a psychiatrist, once told me about one of his patients. This patient was delusional: he thought that he had a squirrel in his head. Odd delusions of this nature do occur, and this patient was adamant about the squirrel. When told that a cranial rodent was illogical and incompatible with physics, he agreed, but then went on to note that logic and physics cannot account for everything in the universe. When asked whether he could feel the squirrel — that is to say, whether he suffered from a sensory hallucination — he denied any particular feeling about it. He simply knew that he had a squirrel in his head.

We can ask two types of questions. The first is rather foolish but I will spell it out here. How does that man's brain produce an actual squirrel? How can neurons secrete the claws and the tail? Why doesn't the squirrel show up on an MRI scan? Does the squirrel belong to a different, non-physical world that can't be measured with scientific equipment? This line of thought is, of course, nonsensical. It has no answer because it is incoherent.

The second type of question goes something like this. How does that man's brain process information so as to attribute a squirrel to his head? What brain regions are involved in the computations? What history led to that strange informational model? Is it entirely pathological or does it in fact do something useful?

So far, most brain-based theories of consciousness have focused on the first type of question. How do neurons produce a magic internal experience? How does the magic emerge from the neurons? The theory that I am proposing dispenses with all of that. It concerns itself instead with the second type of question: how, and for what survival advantage, does a brain attribute subjective experience to itself? This question is scientifically approachable, and the attention schema theory supplies the outlines of an answer.

Attention is a data-handling method used by neurons. It isn't a substance and it doesn't flow

One way to think about the relationship between brain and consciousness is to break it down into two mysteries. I call them Arrow A and Arrow B. Arrow A is the mysterious route from neurons to consciousness. If I am looking at a blue sky, my brain doesn't merely register blue as if I were a wavelength detector from Radio Shack. I am *aware* of the blue. Did my neurons create that feeling?

Arrow B is the mysterious route from consciousness back to the neurons. Arrow B attracts much less scholarly attention than Arrow A, but it is just as important. The most basic, measurable, quantifiable truth about consciousness is simply this: we humans can say that we have it. We can *conclude* that we have it, couch that conclusion into language and then report it to someone else. Speech is controlled by muscles, which are controlled by neurons. Whatever consciousness is, it must have a specific,

physical effect on neurons, or else we wouldn't be able to communicate anything about it. Consciousness cannot be what is sometimes called an epiphenomenon — a floating side-product with no physical consequences — or else I wouldn't have been able to write this article about it.

Any workable theory of consciousness must be able to account for both Arrow A and Arrow B. Most accounts, however, fail miserably at both. Suppose that consciousness is a non-physical feeling, an aura, an inner essence that arises somehow from a brain or from a special circuit in the brain. The 'emergent consciousness' theory is the most common assumption in the literature. But how does a brain produce the emergent, non-physical essence? And even more puzzling, once you have that essence, how can it physically alter the behaviour of neurons, such that you can say that you have it? 'Emergent consciousness' theories generally stake everything on Arrow A and ignore Arrow B completely.

The attention schema theory does not suffer from these difficulties. It can handle both Arrow A and Arrow B. Consciousness isn't a non-physical feeling that emerges. Instead, dedicated systems in the brain compute information. Cognitive machinery can access that information, formulate it as speech, and then report it. When a brain reports that it is conscious, it is reporting specific information computed within it. It can, after all, only report the information available to it. In short, Arrow A and Arrow B remain squarely in the domain of signal-processing. There is no need for anything to be transmuted into ghost material, thought about, and then transmuted back to the world of cause and effect.

Some people might feel disturbed by the attention schema theory. It says that awareness is not something magical that emerges from the functioning of the brain. When you look at the colour blue, for example, your brain doesn't generate a subjective experience of blue. Instead, it acts as a computational device. It computes a description, then attributes an experience of blue to itself. The process is all descriptions and conclusions and computations. Subjective experience, in the theory, is something like a myth that the brain tells itself. The brain *insists* that it has subjective experience because, when it accesses its inner data, it finds that information.

I admit that the theory does not feel satisfying; but a theory does not need to be satisfying to be true. And indeed, the theory might be able to explain a few other common myths that brains tell themselves. What about out-of-body experiences? The belief that awareness can emanate from a person's eyes and touch someone else? That you can push on objects with your mind? That the soul lives on after the death of the body? One of the more interesting aspects of the attention schema theory is that it does not need to turn its back on such persistent beliefs. It might even explain their origin.

The heart of the theory, remember, is that awareness is a model of attention, like the general's model of his army laid out on a map. The real army isn't made of plastic, of course. It isn't quite so small, and has rather more moving parts. In these respects, the model is totally unrealistic. And yet, without such simplifications, it would be impractical to use.

If awareness is a model of attention, how is it simplified? How is it inaccurate? Well, one easy way to keep track of attention is to give it a spatial structure — to treat it like a substance that flows from a source to a target. In reality, attention is a data-handling method used by neurons. It isn't a substance and it doesn't flow. But it is a neat accounting trick to model attention in that way; it helps to keep track of who is attending to what. And so the intuition of ghost material — of ectoplasm, mind stuff that is generated inside us, that flows out of the eyes and makes contact with things in the world — makes some sense. Science commonly regards ghost-ish intuitions to be the result of ignorance, superstition, or faulty intelligence. In the attention schema theory, however, they are not simply ignorant mistakes. Those intuitions are ubiquitous among cultures because we humans come equipped with a handy, simplified model of attention. That model informs our intuitions.

Many people believe that they can feel a subtle heat when someone is staring at them

What are out-of-body experiences then? One view might be that no such things exist, that charlatans invented them to fool us. Yet such experiences can be induced in the lab, as a number of scientists have now shown. A person can genuinely be made to feel that her centre of awareness is disconnected from her body. The very existence of the out-of-body experience suggests that awareness is a computation and that the computation can be disrupted. Systems in the brain not only compute the information that I am aware, but also compute a spatial framework for it, a location, and a perspective. Screw up the computations, and I screw up my understanding of my own awareness.

And here is yet another example: why do so many people believe that we see by means of rays that *come out* of the eyes? The optical principle of vision is well understood and is taught in elementary school. Nevertheless, developmental psychologists have known for decades that children have a predisposition to the opposite idea, the so-called 'extramission theory' of vision. And not only children: a study by the psychologist Gerald Winer and colleagues at the University of Ohio in 2002 found that about half of American college students also think that we see because of rays that come out of the eyes. Our culture, too, is riddled with the extramission theory. Superman has X-ray vision that emanates from his eyes toward objects. The Terminator has red glowing eyes. Many people believe that they can feel a subtle heat when someone is staring at them. Why should a physically inaccurate description of vision be so persistent? Perhaps because the brain constructs a simplified, handy model of attention in which there is such a thing as awareness, an invisible, intangible stuff that flows from inside a person out to some target object. We come pre-equipped with that intuition, not because it is physically accurate but because it is a useful model.

Many of our superstitions — our beliefs in souls and spirits and mental magic — might emerge naturally from the simplifications and shortcuts the brain takes when representing itself and its world. This is not to say that humans are necessarily trapped in a set of false beliefs. We are not forced by the built-in wiring of the brain to be superstitious, because there remains a distinction between intuition and intellectual belief. In the case of ventriloquism, you might have an unavoidable gut feeling that consciousness is emanating from the puppet's head, but you can still understand that the puppet is in fact inanimate. We have the ability to rise above our immediate intuitions and predispositions.

Let's turn now to a final — alleged — myth. One of the long-standing questions about consciousness is whether it really does anything. Is it merely an epiphenomenon, floating uselessly in our heads like the heat that rises up from the circuitry of a computer? Most of us intuitively understand it to be an active thing: it helps us to decide what to do and when. And yet, at least some of the scientific work on consciousness has proposed the opposite, counter-intuitive view: that it doesn't really *do* anything at all; that it is the brain's after-the-fact story to explain itself. We act reflexively and then make up a rationalisation.

There is some evidence for this post-hoc notion. In countless psychology experiments, people are secretly manipulated into making certain choices — picking green over red, pointing left instead of right. When asked why they made the choice, they confabulate. They make up reasons that have nothing to do with the truth, known only to the experimenter, and they express great confidence in their bogus explanations. It seems, therefore, that at least some of our conscious choices are rationalisations after the fact. But if consciousness is a story we tell ourselves, why do we need it? Why are we aware of anything at all? Why not just be skilful automata, without the overlay of subjectivity? Some philosophers think we *are* automata and just don't know it.

This idea that consciousness has no leverage in the world, that it's just a rationalisation to make us feel better about ourselves, is terribly bleak. It runs against most people's intuitions. Some people might confuse the attention schema theory with that nihilistic view. But the theory is almost exactly the opposite. It is not a theory about the uselessness or non-being of consciousness, but about its central importance. Why did an *awareness of stuff* evolve in the first place? Because it had a practical benefit. The purpose of the general's plastic model army is to help direct the real troops. Likewise, according to the theory, the function of awareness is to model one's own attentional focus and control one's behaviour. In this respect, the attention schema theory is in agreement with the common intuition: consciousness plays an active role in guiding our behaviour. It is not merely an aura that floats uselessly in our heads. It is a part of the executive control system.

In fact, the theory suggests that even more crucial and complex functions of consciousness emerged through evolution, and that they are especially well-developed in humans. To attribute awareness to oneself, to have that computational ability, is the first step towards attributing it to others. That, in turn, leads to a remarkable evolutionary transition to social intelligence. We live embedded in a matrix of perceived consciousness. Most people experience a world crowded with other minds, constantly thinking and feeling and choosing. We intuit what might be going on inside those other minds. This allows us to work together: it gives us our culture and meaning, and makes us successful as a species. We are not, despite certain appearances, trapped alone inside our own heads.

And so, whether or not the attention schema theory turns out to be the correct scientific formulation, a successful account of consciousness will have to tell us more than how brains become aware. It will also have to show us how awareness changes us, shapes our behaviour, interconnects us, and makes us human.

Published on 21 August 2013