NEUROSCIENCE of Consciousness

Neuroscience of Consciousness, 2022, **2022(1)**, 1–9

DOI: https://doi.org/10.1093/nc/niac001

Research Article

Special Issue: Consciousness science and its theories

Consciousness explained or described?

Aaron Schurger[1,2,3,4,†,*] and Michael Graziano[5,6]

[1]Department of Psychology, Crean College of Health and Behavioral Sciences, Chapman University, One University Drive, Orange, CA 92867, USA; [2]Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University, 14725 Alton Pkwy, Irvine, CA 92618, USA; [3]Cognitive Neuroimaging Unit, NeuroSpin center, INSERM, Gif sur Yvette 91191, France; [4]Commissariat à l'Energie Atomique, Direction des Sciences du Vivant, I2BM, NeuroSpin Center, Gif sur Yvette 91191, France; [5]Department of Psychology, Princeton University, Princeton, NJ 08540, USA;[6]Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540, USA

[†]Aaron Schurger, http://orcid.org/0000-0003-2985-3253

*Correspondence address. Department of Psychology, Institute for Interdisciplinary Brain and Behavioral Sciences, Chapman University, One University Drive, Orange, CA 92867, USA. E-mail: schurger@chapman.edu

Abstract

Consciousness is an unusual phenomenon to study scientifically. It is defined as a subjective, first-person phenomenon, and science is an objective, third-person endeavor. This misalignment between the means—science—and the end—explaining consciousness—gave rise to what has become a productive workaround: the search for 'neural correlates of consciousness' (NCCs). Science can sidestep trying to explain consciousness and instead focus on characterizing the kind(s) of neural activity that are reliably correlated with consciousness. However, while we have learned a lot about consciousness in the bargain, the NCC approach was not originally intended as the foundation for a true explanation of consciousness. Indeed, it was proposed precisely to sidestep the, arguably futile, attempt to find one. So how can an account, couched in terms of neural correlates, do the work that a theory is supposed to do: explain consciousness? The answer is that it cannot, and in fact most modern accounts of consciousness do not pretend to. Thus, here, we challenge whether or not any modern accounts of consciousness are in fact theories at all. Instead we argue that they are (competing) laws of consciousness. They describe what they cannot explain, just as Newton described gravity long before a true explanation was ever offered. We lay out our argument using a variety of modern accounts as examples and go on to argue that at least one modern account of consciousness, attention schema theory, goes beyond describing consciousness-related brain activity and qualifies as an explanatory theory.

Keywords: theory; law; attention schema theory; HOTT; GNWT; IIT

It seems so obvious that the Sun goes around the Earth—just look up in the sky. From most vantage points, the sun rises in the east, crosses the sky, and sets in the west every 24 hours without fail. And we certainly do not feel the Earth moving beneath our feet. These observations are undeniable and yet we know them to be misleading, thanks in large measure to Galileo who championed Copernicus' heliocentric theory: The Earth revolves around the Sun and not vice versa. And the Earth rotates on its axis every 24 hours, which gives the impression that the sun goes around the Earth. Importantly, Galileo not only explained the phenomenon but also explained why we might be inclined to believe otherwise. The Copernican heliocentric theory is a true theory: it does not just describe what appears in the sky but also explains why it appears that way.

Contrast Galileo's theory with Newton's law of gravity. Newton tells us to accept the truth of the equation $F = G\frac{m_1 m_2}{r^2}$. Never mind why, it just is. This is not a theory. It may be correct. It may lead to countless correct predictions, but still it is not a theory because it does not explain anything. It just describes things. Famously, Newton's law of gravity left out any account of what could possibly cross from one mass to another and exert a force. It was not until Einstein that we had a true theory of gravity—an explanation. Einstein suggested that space-time itself can be curved and that changes in curvature transmit like a wave at the speed of light. Note that both provide a means to make predictions, although Einstein's version makes more accurate and novel predictions.

Theory versus law in science

Simply put, and in the context of this article, a theory is an explanation, whereas a law is a description. Both can be true or false, and, importantly, both can lead to correct (or incorrect) predictions. According to Webster's dictionary, a theory is 'a plausible or scientifically acceptable general principle or body of principles

Received: 8 February 2021; **Revised:** 23 December 2021; **Accepted:** 5 January 2022

offered to *explain* phenomena' (emphasis added; every source that we could find includes the notion of explanation). A law, by contrast, is 'a statement of an order or relation of phenomena that, so far as is known, is invariable under the given conditions'. A law describes the way things work without giving an explanation. Both laws and theories can lead to correct predictions, but a true theory goes further. It offers an explanation. These will be our working definitions of theory and law in the context of this article. There is a large body of work in the philosophy of science and philosophy of mind on the difference between a theory and a law and what constitutes an explanation (Boyd and Bogen 2021; Woodward and Ross 2021). Our goal here is not to enter into that longstanding debate but rather to start with the definitions offered above, couched in terms of explanation and description and work from there.

Is this just semantics? Well, no, not any more than the distinction between a description and an explanation is merely semantic. We could potentially come up with a perfectly accurate description of what goes on in the brain that 'gives rise to' consciousness without having any clue as to why. That is the gist of Levine's explanatory gap (Levine 1983). The idea of the explanatory gap is that, while it is possible to come up with laws of consciousness, a true scientific theory of consciousness is not possible (Goff 2019) (assuming you identify 'consciousness' with qualia or subjective experience). Here, we take up the argument that most modern theories of consciousness, even the most prominent ones, are not theories at all but rather are (proposed) laws. Some of them can feasibly be tested and could end up being right or wrong, but, ultimately, they are not theories because they only describe a relation but do not offer an explanation.

We conclude with a discussion of a relative newcomer on the theoretical landscape, attention schema theory (AST) (Graziano and Kastner 2011). We argue that AST is in fact a theory because it tries to explain something, even if that something is not qualia *per se* but rather the perception (or belief) that we have qualia in the first place as something above and beyond neural information combined with selective attention. AST might end up being supported or refuted by evidence—it could be right or wrong—but it meets the criteria for being a theory. AST gets away with being a theory because it does not treat consciousness as something inherently mysterious, ineffable, irreducible, or unexplainable using a reductionist or mechanistic framework (Levine 1983; Chalmers 1996; Searle 1997; Tononi *et al.* 2016; Koch 2018). AST's explanatory power may in part be owing to the way it approaches the problem of consciousness—by offering to explain why we believe in consciousness rather than explaining consciousness *per se*—and many people find this unsatisfactory. As a consequence, some might claim that AST does not explain 'consciousness' at all—maybe it explains our beliefs about consciousness, or our introspective certainty that we have it (Frankish 2016) but not 'consciousness' itself—*and they would be correct*, if by 'consciousness' they mean qualia. This observation exposes a somewhat neglected front in the debate about the nature of consciousness: *What is it that we are trying to explain? What is it that we should be trying to explain? What is the minimal set of facts that we need to account for in order for our explanation to be complete?*

Three theories of consciousness that are really proposed laws

There are many competing accounts of consciousness that claim to be theories. They have been cataloged, compared, and contrasted elsewhere (Doerig *et al.* 2020). Many scholars, like Chalmers (Chalmers 1996), have argued that no scientific theory can truly explain consciousness. This is what Leibniz tried to convey in his analogy of the mill (Monadology 1714; see Discussion). It means that when we try to explain consciousness, we arrive at a chasm where we say 'and then consciousness happens' in order to magically jump over the explanatory gap. Virtually all accounts of consciousness have this in common. They walk you to the edge (from different angles) and then somehow you find yourself on the other side, without understanding how you got there. They do not solve the hard problem, but, in all fairness, most do not claim to either. Let us consider three popular neuroscientific accounts of consciousness as examples: global neuronal workspace theory (GNWT), integrated information theory (IIT), and higher-order thought theory (HOTT).

According to GNWT (Mashour *et al.* 2020), neural information becomes conscious when the neural activity carrying that information gains access, in a winner-take-all fashion, to the global neuronal workspace. The global workspace is a widely distributed network of long-range connections originating in parietal and prefrontal cortices. When neural information enters, or takes command of, the global workspace, then that information can be *broadcast* widely throughout the brain, allowing for that information to be globally accessible to multiple specialized centers in the brain involved in attention, valuation, memory, perception, motor output, and so on. One key feature of GNWT is that the global workspace can only entertain one interpretation of the sensory/perceptual data at a time such that when the global workspace settles on a given interpretation, then other competing interpretations are temporarily blocked out of conscious experience. Therefore, the workspace, and, by extension, consciousness, is a kind of bottleneck from the point of view of information processing. This aspect of the theory has been used to explain phenomena like the attentional blink (Sergent *et al.* 2005) and bi-stable perceptual phenomena like the Necker cube or face-vase illusion.

One might argue, based on the above, that GNWT *does* explain something—it explains why there is a lapse in sensory information processing during the attentional blink, and why the observer can only entertain one interpretation of the Necker cube at a time. But that would be a red herring, vis-à-vis the phenomenon that the theory was supposed to explain—subjective experience. The attentional blink is an information processing phenomenon that happens to have an effect on conscious visual perception, and GNWT does a good job of explaining the attentional blink. But explaining the attentional blink is not equivalent to explaining consciousness. The attentional blink is useful as a paradigm case (Doerig *et al.* 2020) for observing and describing what happens differently in the brain when an otherwise visible visual stimulus is not consciously perceived (compared to when it is). But explaining the attentional blink does not entail explaining *why* global broadcasting of information, via the global workspace, should give rise to or be accompanied by qualia. Note that blinking your eyes also has a profound effect on conscious visual perception, but no one would assert that an explanation of eye blinking should also constitute an explanation of visual consciousness. Note also that lawful relations, like Newton's law of gravity, can make accurate and useful predictions that can be tested, and this is true of accounts of consciousness like GNWT. A theory goes farther—it explains why.

One way to think about how to effectively test a theory is by considering the selectivity of the test (Merikle *et al.* 2001). Finding a selective test applicable to consciousness is difficult. One could reasonably argue that *there exists no phenomenon that selectively abolishes consciousness and has no impact on anything else*. Well, what

about blindsight (Stoerig and Cowey 1997; Cowey 2004), you ask? Not even close. Guessing ability in blindsight is severely limited compared to normal conscious vision. A hemianopic patient with blindsight might be able to correctly guess the orientation of a bar significantly more often than chance (Schurger *et al.* 2006) but then not be able to discriminate between a square and a rectangle (Stoerig and Cowey 1997; Cowey 2004) and be utterly unable to identify the category of visual objects (like 'tree', 'car', 'house', and 'cow'). And, in the case of Type I blindsight (Cowey 2004), if you do not cue the patient to produce a guess, then the patient may not respond at all. But even if there were a manipulation that selectively abolished e.g. conscious visual perception and nothing else, it could only serve to make the description, or lawful relation, more precise. A mechanistic explanation of that manipulation, no matter how precise, would still not constitute an explanation of subjective experience. In the case of GNWT, the step from 'entry into the global workspace' to 'qualia' is a leap of faith.

As another example, consider IIT (Balduzzi and Tononi 2008). According to IIT, consciousness is integrated information. To the extent that information is at once integrated and differentiated, that information is conscious. Integration and differentiation are considered to be axiomatic properties of consciousness, asserted without proof or empirical evidence other than an appeal to introspection (which is taken to provide ground truths). As with Newton's law of gravity, IIT can be expressed in the form of an equation by which the degree of information integration can be computed and assigned a value called ϕ (phi). A system with zero ϕ is not conscious, and systems with nonzero ϕ are conscious to varying degrees, with the value of ϕ expressing the degree to which the system is conscious. If ϕ is taken at face value, then according to IIT, some very simple systems that *prima facie* are not conscious are in fact highly conscious because they have a rather high value of ϕ (Aaronson 2014; Horgan 2015). This is not a problem for IIT, however, since the theory openly embraces pan-psychism—the idea that consciousness is a fundamental property of physical systems, much like inertia, and that everything in the universe, with nonzero ϕ, possesses it to some degree. Thus, rocks and trees and the solar system may have a degree of consciousness. Unlike with inertia, however, you cannot reach out and verify the presence of consciousness in e.g. a sea slug. You could compute ϕ in the sea slug, but then you would have to accept the assertion that ϕ selectively indexes consciousness and not something else that happens to covary with consciousness in humans. Although (highly) impractical, ϕ can in theory be computed for systems like brains but not without computing power that we can hardly envision, let alone realize.

IIT does not have parsimony on its side—it is a very complex theory, and the math behind computing ϕ is orders of magnitude more complex than the math behind Newton's law of gravitation. In fact, an early exposition of IIT (before it was given that moniker) equated consciousness with complexity (Tononi and Edelman 1998). But we do not need to delve into that complexity in order to evaluate whether IIT explains or merely describes consciousness. The equation for ϕ, much like Newton's mathematical formulation of his law of gravity, describes a lawful relation, but IIT does not explain why that lawful relation holds. The step from ϕ to subjective experience is a leap of faith. IIT describes something that is alleged to covary with consciousness (the result of a mathematical equation and thus not identical to consciousness), but it does not explain why qualia are present when $\phi > 0$ and absent otherwise.

IIT makes some predictions about how the complexity of brain activity should be a reliable signature of consciousness (Massimini *et al.* 2009) and this, in turn, has contributed to the development of a technique for inferring the presence or absence of consciousness in noncommunicating patients with disorders of consciousness (Casali *et al.* 2013). The same metric may also have prognostic value in predicting the probability of recovery of consciousness (Rosanova *et al.* 2012). Does not this speak in favor of IIT being a true explanatory theory? Not by any means, no. First of all, no reported case of brain damage, through injury or stroke, has ever selectively abolished only subjective experience, leaving all other brain functions intact. Any such case study would be famous beyond all reckoning, in every textbook, and known to all of us in the field (were it not for the fact that it would be impossible to diagnose[1]). Therefore, any brain function that is intimately related to, but not identical to, consciousness, like selective attention, working memory, or perception, could be used as a reliable signature of consciousness. IIT might be a superb account of perception, and not a theory of consciousness, and yet still lead to the development of a highly accurate consciousness meter because brain damage tends to affect both perception and subjective experience in roughly equal measure.

As long as there are other neuro-cognitive phenomena that are tightly coupled with consciousness, you have a problem when you rest your case on evidence of an accurate consciousness meter. Your consciousness meter might be detecting one of those other phenomena that are tightly coupled with consciousness and not consciousness itself. If you are going to rest your case on that kind of evidence, then you need disruptions that are highly selective, and brain injuries and strokes almost never are. So, we could completely sideline consciousness, from a theoretical standpoint, and still end up with a good consciousness meter. But the most important reason that inspiring a useful metric does not entail being an explanatory theory is simply that laws can also yield predictions, can be tested, and can lead to useful metrics without necessarily explaining the phenomena that they describe. Newton's law of gravitation declares that $F = G\frac{m_1 m_2}{r^2}$ without explaining why and, yet, it makes perfectly good predictions about the gravitational force between two masses.

As one final example, consider HOTT (Rosenthal 2005). According to HOTT, a brain state is conscious to the extent that one is representing oneself as being in that state. Therefore, conscious states are states about which one has formed a higher-order thought or higher-order representation. Originally HOTT was not a neuroscientific theory but later took on a neuroscientific dimension when theorists began suggesting that specific regions of the brain, namely the prefrontal cortex (PFC), were responsible for instantiating higher-order representations (Lau and Rosenthal 2011). Forming higher-order representations might be intimately related to the faculty of metacognition, which is thought to depend on the PFC (Fleming *et al.* 2010). An important caveat here is that these two ideas, rather than being two facets of the same theory, are in fact more or less orthogonal. One says that HOTs are necessary and sufficient for consciousness,

[1] We have essentially described a philosophical zombie in this passage, which is an oft-cited thought experiment that highlights why there is a hard problem in the first place. How would you diagnose such a syndrome? Any question asked of the patient and any input–output test you could conceive of would yield identical results. So, you would have to rely on some measure of brain activity that you had previously associated with conscious processing. Supposing that test came up negative for consciousness? Would you consider the patient to be the first living zombie, or would you question the validity of your measure of consciousness? There is no clear way to adjudicate.

and the other says that PFC is responsible for instantiating HOTs. One could be right while the other is wrong or vice versa, or both could be right, or both could be wrong. But whether we think of them together or separately, neither explains why HOTs (whether instantiated in PFC or elsewhere) should give rise to subjective experience. It may well turn out to be true that wherever there is a HOT there is a conscious state, and vice versa, but we are offered no explanation as to why that should be the case. HOTT is a description (perhaps accurate, perhaps not) of which kinds of states are conscious and which are not.

You might argue that theories of consciousness are not trying to explain consciousness *per se*, as in qualia or subjective experience, but rather are trying to account for the mechanics of consciousness in the brain. This would be the approach inherited from the 'neural correlates of consciousness' (NCC) research program (Crick and Koch 1998). But then what would such an account be trying to explain? What would be the target phenomenon, if not subjective experience? All modern accounts of consciousness are of the form 'when conditions X, Y, and Z are met then consciousness happens, or is allowed to happen'. This is a description, not an explanation, because it does not address the questions of 'how' and 'why'. Why should it be that when Conditions X, Y, and Z are met, consciousness happens or emerges? This is the very question that the NCC program of research was aimed at skirting, and this was a wise workaround because it enabled empirical research to continue to flourish without getting hung up on a potentially unanswerable question. What was never acknowledged, however, was that, in doing so, we forego the possibility of building a true theory. Essentially, the NCC program of research says, forget about explaining consciousness for now. We may or may not ever be able to do that. But we can describe what goes on in the brain when consciousness is present versus absent. We do not have to explain how consciousness comes to be (i.e. a theory) in order to describe the neural correlates of consciousness (i.e. a law).

In all fairness to the accounts of consciousness discussed above, and to the many other accounts that we have not discussed, none claims to solve the hard problem of consciousness. By asserting that they do not, we are neither condemning nor even criticizing them. We are, however, asserting that they are not true theories of consciousness but rather ought to be thought of as (competing) laws of consciousness. They describe the kind of states or nature of brain activity that underlies consciousness without pretending to explain why.

One might counter that, in the case of some accounts of consciousness, such as IIT, the claim is of an identity relation—like 'water is H2O' or 'lightening is an electrical discharge'—and identities do not need to be explained. But this claim is problematic because, unlike other phenomena, we have no direct third-person empirical data about consciousness *per se*, whereas we do (or at least can) with other phenomena like water or lightening. These data are necessary in order to support the claim of an identity relation. Proponents of IIT would argue that we do, however, have direct *first-person* introspective access to properties of consciousness and that those properties can be used to establish the identity relation (Tononi *et al.* 2016). But do we have direct introspective access to the properties of consciousness?

It is easy to confuse 'access to consciousness' with 'access to the contents of consciousness'. One could argue that, in fact, what we really have direct introspective access to are *the contents of consciousness* and not consciousness *per se*. For example, when one declares one's own consciousness to be unified, another could counter by asserting that it is *the contents of consciousness* (i.e. neural information) that are judged to be unified.[2] Normally, when we consciously perceive a perceptual object—a face, for example—we can immediately report on the properties of that information content, but we have no introspective access whatsoever as to how that information content came to be conscious—which is the central question in the scientific study of consciousness. In fact, consciousness *per se* is arguably one phenomenon that we do not have direct introspective access to. Claiming that one has 'direct undeniable first-person introspective access to the properties of consciousness' thus entails a commitment to the assertion that perception (i.e. the formation of internal representations in the form of neural activity) is identical to consciousness. But if that were the case, then we would not need a science of consciousness in the first place.

Attention schema theory

We argue here that at least one proposed theory of consciousness, AST, is indeed an explanatory theory and not an asserted law. It does not merely describe the conditions in which consciousness occurs but explains how those conditions result in (among other things) the objectively measurable phenomenon of people being absolutely certain that they have a conscious experience. Some might argue that comparing AST to other accounts of consciousness is vacuous and ultimately misleading because, in contrast to other accounts, AST is not trying to explain subjective experience *per se* but rather is trying to explain the belief in and beliefs about subjective experience—the explanandum is not the same. While true that the explanandum is indeed not the same, this fact does not grant a license to dismiss AST as irrelevant to consciousness because AST is not simply a theory of why people believe themselves to have qualia. That would be akin to claiming that Einstein's theory of gravity is simply a theory of why apples fall from trees to the ground. In fact, it is a theory of something called gravity, and that theory happens to explain, among other things, why apples fall from trees to the ground. Likewise, AST is a theory of something called consciousness that explains why, among other things, people believe themselves to have qualia. Consciousness, according to AST, is a special kind of percept that arises due to the workings of a hypothetical mechanism called an 'attention schema'. The attention schema helps to guide, stabilize, and control selective attention, and having an attention scheme can lead to an adamant belief in an ineffable something extra that we might call qualia. This necessarily implies a different sort of explanandum from the one we typically associate with the mystery of consciousness, but this is a feature of AST, not a bug. To see why, consider the following story:

Suppose that a community of ancient scholars debates a question. The explanandum that interests them: at the end of each day, the sun sets down somewhere in the distance to the west and its fire is extinguished. And each morning it is reignited in the east and travels across the sky throughout the day only to be extinguished once again in the west. How does that happen? They propose a host of ideas, perhaps some of them true theories (attempts to explain) and some of them laws (descriptions of the circumstances). Now imagine that one of those scholars, let us call him Jake, comes up with a new idea: The sun does not set down or go out. Rather, the phenomenon to explain is the movement of heavenly bodies around each other. We cannot see the

[2] And even that is debatable: there may be clinical cases, such as in Balint's syndrome, where the patient might declare that consciousness is not unified—even though we have no reason to doubt that the patient is in fact conscious. This is an empirical question and has yet, to our knowledge, been tested.

sun at night because, as the earth rotates, the bulk of the earth stands between it and us. The ancient scholars could accuse Jake of cheating, 'solving' the problem simply by denying the original explanandum—cutting the Gordian knot instead of untying it. They may be right, from one philosophical point of view. But Jake is still correct. With a re-conceptualization of the problem, and a new explanandum, the problem makes more sense, and an explanatory theory becomes possible. It is absolutely fair to compare Jake's explanation to the prevailing explanation, even if they have different explananda. Scientific progress depends on making such comparisons and tossing out the conceptual framework that fails to produce a viable and veritable theory.

In addition to the argument above, we also remind the reader that we are making two different claims in our article, and only the latter of the two is subject to the criticism about explananda. Our main claim (elucidated earlier) is that most purported theories of consciousness are not in fact theories. Our argument in support of that claim is agnostic as to what the target phenomenon (or explanandum) is. Whatever the explanatory target happens to be—an extra something, or the belief in an extra something—either way, a theory must do more than just describe the physical facts that reliably accompany the target phenomenon. We argue that most current accounts of consciousness are not in fact theories *of that which they claim to be theories of*.

Our second claim (argued below) is about what AST in particular is trying to explain and how AST is in fact an explanatory theory. AST explains why people are convinced that they have this extra something. AST does not explain how this extra something comes about. One could assert that, according to AST, consciousness simply does not exist. We only believe it does, and it is an illusion. This characterization may be true, but it misses much of the meaning of AST. The whole point of AST is to link the belief that we have consciousness with an actual thing that we actually have: selective attention. Consciousness is not an empty illusion. It is a distortion or caricature of something real: selective attention. AST explains how people can control their attention as well as they do (because they have an internal model of it), and AST explains how people can predict the behavior of others so well (because we construct models of their attention, and what they attend to tends to drive their behavior), and AST also explains why people believe, insist, think, and claim to have something—subjective experience—that defies any kind of reductionist or mechanistic explanation. Yes, AST explains why we believe we have consciousness, but to insist only on that facet of AST misses the bigger picture. It is a bit like trying to explain how an automobile engine works to someone who is primarily interested in the phenomenon of the sound coming from the engine. Yes, AST can explain that sound, but that is barely scratching the surface of what AST can explain as an engineering theory of how to build and understand the engine.

AST is a specific version of what might be called the self-model approach to consciousness. 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—must stem from information[3] in the brain, or you would not be able to think the thought or articulate the claim. What we believe to be true depends on bundles of information, effectively models. The brain is a model builder. The visual system builds visual models, rich sets of information that represent the shape and color and movement of objects. The body schema is a set of information about kinematic, dynamic, and structural aspects of the body. Our intellectual beliefs about the world and political, religious, and scientific beliefs are models at a more cognitive level. But, as the statistician George Box is supposed to have said, 'All models are wrong; some are useful'. The brain's models are always approximate rather than literally accurate. If the brain built fully accurate, detailed models of everything relevant to survival, it would run out of processing power posthaste.

Now, we can approach the question: why are people so convinced that they have an ineffable, subjective feel that accompanies their thinking and their perception? Why are people so convinced that they have consciousness? Chalmers has called this question the 'meta-problem' of consciousness (Chalmers 2018). Logically, the answer is that the human brain constructs an information set, a part of a self-model, on the basis of which people derive the belief and certainty that they have a conscious experience. Moreover, if this self-model is like every other model studied in neuroscience or psychology, then it is likely to be an inexact representation—a simplification and distortion of some actual, physically measurable property of the self.

Thus far, the self-model view may seem odd. After all, we do not *believe* we have consciousness. We simply have it. We are used to the intuitively compelling argument that 'I know I have a conscious feeling inside me, because I'm experiencing it right now. I can feel it'. But this argument is a tautology. It is equivalent to saying, 'I know it's true because it's true. I know I have a feeling, because I *feel* the feeling'. When the architects of IIT assert that the presence of consciousness is axiomatically true or, as Tononi puts it, the one thing that we know for certain, they are engaging in tautology. Instead, what we know with some reasonable scientific certainty is that the brain has constructed a set of information on the basis of which a belief is derived. Cognition has gained access to an information set; the information set is part of a self-model; based on that information, cognition arrives at the belief and the certainty that a conscious feeling is present. We think we have something nonphysical and intangible inside us, a hard problem, 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 self-model perspective has deep roots in a philosophical view called illusionism (Dennett 1988, 1991; Frankish 2016). The term illusionism is, however, arguably not the best label. Perhaps, rather than suggesting that consciousness is an illusion, it would be more apt to say that it is a perceptual caricature. Unlike the term illusion, the term caricature implies the presence of something real that is being caricatured. In this self-model perspective, something physically real and objectively measurable exists inside us; the brain constructs a simplified, distorted model of that physical mechanism; and on the basis of that model, we believe we have an essentially magical, conscious experience. The scientific question of consciousness then becomes: what is the real mechanism that gives rise to the self-model on which our belief in a hard problem of consciousness depends? AST is a specific theory that addresses that question.

AST relates consciousness to attention (Graziano and Kastner 2011; Graziano 2013; Webb and Graziano 2015). It proposes that selective attention is a real, physical process in the brain; the brain constructs a simplified, descriptive model of attention, termed the attention schema (in parallel to the body schema); and on the basis of the information in the attention schema, higher cognition arrives at the belief and certainty that a subjective conscious experience is present. One can arrive at the theory from several

[3] When we refer to information, we mean the resolution of uncertainty or an arrangement of matter and/or energy that can resolve uncertainty.

directions, but one especially obvious path starts with the close relationship between consciousness and 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 (Desimone and Duncan 1995; Corbetta and Shulman 2002; Beck and Kastner 2009; Moore and Zirnsak 2017). The brain contains control mechanisms that allow it to strategically shift attention not only among sensory events but also to shift focus among memories and thoughts. Attention, or a controlled and selective signal enhancement, is an entirely mechanistic, physical process that can be measured objectively by a variety of means, some involving direct measurement of neuronal events in the brain and some involving the measurement of behavioral accuracy and latency.

Consciousness is similar in many ways to attention. When we are conscious of something, we feel that we are processing it, we grasp it with the mind, and we are able to respond to it. Moreover, attention and consciousness almost always move together. What your brain is attending to, you are almost always subjectively aware of. Nothing shows this close relationship better than the famous gorilla experiment of Simons and Chabris (Simons and Chabris 1999). People who watched a video of a basketball game, focusing their attention on the basketball, were totally unaware of the man in a gorilla suit dancing right across the center of the scene. By keeping attention on Object A and away from B, people attached their subjective consciousness to A and were unaware of B.

Yet, consciousness and attention can be separated. In laboratory circumstances, it is possible for a person's attention to be exogenously drawn to a visual stimulus, while at the same time, the person reports no subjective consciousness of the stimulus (Woodman and Luck 2003; Tsushima *et al.* 2006; Kentridge *et al.* 2008). When attention and consciousness become decoupled, although attention can remain, one aspect of attention is drastically impaired. The control of attention disappears (Webb *et al.* 2016; Wilterson *et al.* 2020). People lose the ability to endogenously suppress attention, sustain attention, or strategically shift attention with respect to the stimulus of which they are unconscious. Thus, although attention can exist without consciousness, it is not quite right to say that attention is independent of consciousness. The two have a complex and intertwined relationship.

AST proposes that this complex relationship between consciousness and attention reflects a simple, underlying mechanism. To help control its own attention, the brain constructs a simplified model of attention. The information in that model, when accessed by higher cognition, leads to the belief that we have subjective consciousness. For example, when you look at an apple, your knowledge about the apple—its color 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—stems from an attention schema, an imprecise but useful model of the process of attending (to the apple, in this case).

Computational models and computer simulations

Some accounts of consciousness, like GNWT, have been implemented in the form of a computational model/computer simulation (Dehaene *et al.* 1998, 2003). But such computational models beg the question of what, precisely, the model should produce as output in order to be considered a model of consciousness rather than a model of, say, attention, or neural information processing (NIP), or of a specific phenomenon like the attentional blink. Computer simulations of the GNWT do a good job of explaining the attentional blink and its attendant bottleneck in NIP. But what outputs of these simulations should we consider to account specifically for consciousness, rather than accounting for NIP, attention, or the attentional blink? The attentional blink is a phenomenon that happens to have an effect on conscious perception of sensory stimuli, but it is not identical with consciousness. So, a theory that explains the attentional blink is not necessarily a theory of consciousness. GNWT stands quite well on its own as a theory of NIP without having to explain how or why qualia are associated with information that reaches the global neuronal workspace.

AST can also be implemented in the form of a computational model (Wilterson and Graziano 2021). The goal is not to explain qualia *per se* but rather to explain the emergence of utterances about qualia or of having perceived a particular stimulus. Note, however, that a theory does not have to be computational in nature in order to explain its target phenomenon. Darwin's theory of evolution by natural selection can be turned into a computational simulation, but it still has just as much explanatory power even without being implemented as a simulation. Indeed, it existed as a theory since long before it was even possible to simulate evolution (with e.g. genetic algorithms). Likewise, AST counts as an explanatory theory even without having been implemented in a computer simulation.

Perhaps closer in spirit to AST, recent higher-order computational models of awareness, including the 'higher-order state space' model (Fleming 2020), and the 'predictive global neuronal workspace' model (Whyte and Smith 2021) offer the ability to simulate higher-order states that might support a predictive self-model. However, neither account explains how that predictive self-model leads to declarations of having consciousness (whether of state or of content). AST explains how a self-model leads to the belief that 'I am conscious', and in that way AST is fundamentally different. Although AST might be thought of as a kind of HOTT, thinking that HOTs *produce* consciousness betrays a misunderstanding of AST. It is not a theory in which having a HOT makes you conscious. Rather *it is a theory in which you have a HOT that tells you that you have consciousness*. Therefore, you think that you have it. The attention schema is a nonverbal model, so believing that you are conscious of an object does not require a verbal declaration. The question answered by AST is why we believe at a nonverbal, intuitive, automatic level that there is such a thing as conscious experience and that we are having one.

AST is a proper theory

To many people, AST may not be a satisfying theory because it does not explain how the brain generates a subjective essence of consciousness. It sidesteps the hard problem. And, yet, whether you believe the theory or find it lacking, it is an explanatory theory, not an asserted law. It attempts to explain the behavior of the system by offering, among other things, an explanation of why people claim to have consciousness. The types of behaviors that the theory tries to explain are e.g. beliefs and utterances of the form 'There is more to me than just input→processing→ output. I have qualia'. Why do we believe we have subjective awareness? That belief stems from a deeper model, a model of attention. Why do we have a model of attention? Because the human brain has the ability to control its focus of attention, and like all control systems, this one benefits from having a model depicting the thing it controls. Why do people believe subjective awareness is a nonphysical

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, nonphysical essence that can seize hold of items and vividly know them.

The theory may have many gaps. For example, what networks in the brain compute this information? What exactly is the informational content of the attention schema? Can it ever be partial, or is it all or nothing? How does an attention schema relate to components of other theories, such as GNWT or HOTT? Some of these questions are being addressed experimentally and may take many years to resolve. Perhaps, in the end, the data will show the theory to be wrong. But for all the specific scientific gaps, AST does not contain the one, the big explanatory gap. There is no leap of faith. AST is not a proposed law that asserts: when Conditions X, Y, and Z occur, then consciousness emerges. It is a theory that explains how the strong belief in consciousness occurs and also explains the complex, experimentally observed relationship between consciousness and attention.

General discussion

Here, we have argued that most supposed theories of consciousness are not really theories at all but rather are competing laws of consciousness. This is because they do not explain *why* the XYZ of the theory (which is different for different theories) necessarily implies subjective experience. Instead, these accounts merely describe the conditions that are supposed to bring about consciousness. In doing so, these accounts can make testable predictions, but that does not elevate them to the level of being a theory because a law can also make testable predictions. These accounts may also explain various related phenomena (like the attentional blink, for example) that have an effect or impact on subjective experience; but explaining these related phenomena is also not the same as explaining subjective experience or qualia. Indeed, we assert that any 'theory' that is confronted with, but does not bridge, the explanatory gap is not a true theory of consciousness, once we agree that the explanandum is subjective experience. Otherwise, that account would be analogous to asserting that heating milk reduces the risk of disease. It may be true, and it might be useful knowledge, but it is not an explanation. Heating milk reduces the risk of disease because the microorganisms in milk that are responsible for disease cannot tolerate the heat. This account answers the questions 'how' and 'why'.

We also argued that at least one candidate theory of consciousness, AST, does offer a genuine explanation. One might argue that AST, like all other accounts, also fails to bridge the explanatory gap, just in a different way. One could argue that the explanandum in AST is not the right one—it is not the ineffable something extra that is not accounted for in a simple input→processing→output model. But, in fact, AST *is* concerned with the explanatory gap, even though it may seem to sidestep it. AST explains why people are prone to think there is an explanatory gap. Like Galileo's heliocentric theory, it not only explains how things really are but also explains why we might be inclined to believe otherwise. So it does not really turn its back on the hard problem. Instead, it takes the human intuition that a hard problem exists and puts that intuition at the heart of a useful cognitive mechanism. It is saying, no, a magic consciousness essence does not exist. But the belief in that magic essence is actually really important because it acts as a rough and ready model of something else that really does exist: selective attention.

To reiterate, one might argue that AST is not even on the same playing field as other accounts because AST is trying to explain the belief in qualia, whereas other accounts are aimed at qualia *per se*. Since these are fundamentally different things (and potentially independent), there is no point in comparing AST to other accounts of consciousness. But this perspective completely misses the point. AST is trying to wrestle the definition of consciousness away from being a 'something else' that, it seems, we cannot possibly explain in reductionist or mechanistic terms. AST wants to redefine what consciousness is. It is not trying to deny the existence of consciousness or explain consciousness away but is just trying to reframe the question and define consciousness in a different way.

As we mentioned in the introduction, our argument exposes a neglected front in the debate about the neural basis of consciousness: What is it that we are trying to explain? What is it that we should be trying to explain? What is the minimal set of facts that we need to account for in order for our explanation to be complete? Consider the example of studying ghosts. A lot of people believe in them, and pretty adamantly. In fact, there are probably far more people in the world who believe in ghosts and souls than there are people who believe in qualia (given that most people in the world do not even know what the word 'qualia' means). Maybe there is something to it, who knows. How might we approach a question like this, scientifically? We could throw everything we have—technology, equipment, theory, math, physics, … everything—at trying to understand how ghosts come about. Or, we could focus our efforts on studying why people believe in the existence of ghosts. That would also be a valid avenue of research, perhaps orthogonal to the original question of whether or not ghosts exist in the first place. The question then is this: if we succeed in coming up with a clear and comprehensive account of why people believe in ghosts, are we done? Is there anything else in need of explaining, given that our best efforts to find and measure ghosts have all come up dry?

The state of affairs with consciousness research is not so different from this caricature. Great thinkers for centuries have been saying essentially the same thing: The mind (aka consciousness) is really perplexing. It sure seems like there must be something extra in there beyond just input→processing→output, yet, try as we may, there is no sign of anything extra—at least not that we can physically measure. Leibniz's Mill, alluded to in the introduction, is a classic example:

> *It must be confessed, moreover, that perception, and that which depends on it, are inexplicable by mechanical causes, that is, by figures and motions. And, supposing that there were a mechanism so constructed as to think, feel and have perception, we might enter it as into a mill. And this granted, we should only find on visiting it, pieces which push one against another, but never anything by which to explain a perception.*
>
> Gottfried Leibniz, *Monadology*, 1714

Take a step back and look at it from a distance: Imagine that zombie aliens come to visit the Earth from another planet, and they learn that some of us share a very adamant belief in something as yet never before physically detected called qualia. With no institutional review board to stand in their way, they rifle around in human brains with sophisticated equipment, and, as Leibniz predicted, they find nothing but neurons and glia and the like. And at the same time, they conclude that input→processing→output is enough to account for, well, pretty much everything else, *including the belief in qualia itself*. What might those visitors from another

world rightfully conclude about this extra something that we call 'qualia'? AST explains that conviction as a necessary by-product of a useful mechanism: an internal model of selective attention. With a viable explanation like this in hand, and Occam's razor as a guiding principle, what should we conclude? As long as consciousness is defined as something intrinsically inaccessible from the outside, then accounts that try to explain it can never be true scientific theories.

Conflict of interest statement

None declared.

References

Aaronson S. 2014 *Why I Am Not An Integrated Information Theorist (or, The Unconscious Expander)*. https://www.scottaaronson.com/blog/?p=1799 (3 January 2022, date last accessed).

Balduzzi D, Tononi G. Integrated information in discrete dynamical systems: motivation and theoretical framework. *PLoS Comput Biol* 2008;**4**:e1000091.

Beck DM, Kastner S. Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Res* 2009;**49**:1154–65.

Boyd NM, Bogen J. Theory and observation in science. In: Zalta EN (ed.), *Stanford Encyclopedia of Philosophy*. Summer 2021 edn. Stanford, CA: Metaphysics Research Lab, Stanford University, 2021.

Casali AG, Gosseries O, Rosanova M *et al.* A theoretically based index of consciousness independent of sensory processing and behavior. *Sci Transl Med* 2013;**5**:198ra105.

Chalmers D. The meta-problem of consciousness. *J Conscious Stud* 2018;**25**:6–61.

Chalmers DJ. *The Conscious Mind: In Search of a Fundamental Theory*. London: Oxford University Press, 1996.

Corbetta M, Shulman GL. Control of goal-directed and stimulus-driven attention in the brain. *Nat Rev Neurosci* 2002;**3**:201–15.

Cowey A. The 30th Sir Frederick Bartlett Lecture. Fact, artefact, and myth about blindsight. *Q J Exp Psychol* 2004;**57A**:577–609.

Crick F, Koch C. Consciousness and neuroscience. *Cereb Cortex* 1998;**8**:97–107.

Dehaene S, Kerszberg M, Changeux JP. A neuronal model of a global workspace in effortful cognitive tasks. *Proc Natl Acad Sci USA* 1998;**95**:14529–34.

Dehaene S, Sergent C, Changeux JP. A neuronal network model linking subjective reports and objective physiological data during conscious perception. *Proc Natl Acad Sci U S A* 2003;**100**:8520–5.

Dennett DC. Quining qualia. In: Marcel AJ, Bisiach E (eds.), *Consciousness in Modern Science*. Oxford: Oxford University Press, 1988.

Dennett DC. *Consciousness Explained*. Boston: Little, Brown, and Co, 1991.

Desimone R, Duncan J. Neural mechanisms of selective visual attention. *Annu Rev Neurosci* 1995;**18**:193–222.

Doerig A, Schurger A, Herzog MH. Hard criteria for empirical theories of consciousness. *Cogn Neurosci* 2020;**12**:41–62.

Fleming SM. Awareness as inference in a higher-order state space. *Neurosci Conscious* 2020;**2020**:niz020.

Fleming SM, Weil RS, Nagy Z *et al.* Relating introspective accuracy to individual differences in brain structure. *Science* 2010;**329**:1541–3.

Frankish K. Illusionism as a theory of consciousness. *J Conscious Stud* 2016;**23**:11–39.

Goff P. *Galileo's Error: Foundations for a New Science of Consciousness*. New York: Pantheon Books, 2019.

Graziano MSA. *Consciousness and the Social Brain*. Oxford, UK: Oxford University Press, 2013.

Graziano MSA, Kastner S. Human consciousness and its relationship to social neuroscience: a novel hypothesis. *Cogn Neurosci* 2011;**2**:98–113.

Horgan J. *Can Integrated Information Theory Explain Consciousness?* Scientific American, 2015. https://blogs.scientificamerican.com/cross-check/can-integrated-information-theory-explain-consciousness/ (3 January 2022, date last accessed).

Kentridge RW, Nijboer TCW, Heywood CA. Attended but unseen: visual attention is not sufficient for visual awareness. *Neuropsychologia* 2008;**46**:864–9.

Koch C. What is consciousness? *Nature* 2018;**557**:S8–12.

Lau H, Rosenthal D. Empirical support for higher-order theories of conscious awareness. *Trends Cogn Sci* 2011;**15**:365–73.

Leibnitz GW, Rescher N. *GW Leibniz's Monadology: An Edition for Students*. Pittsburgh, PA: University of Pittsburgh Press, 1714/1991.

Levine J. Materialism and qualia: the explanatory gap. *Pacific Phil Quart* 1983;**64**:354–61.

Mashour GA, Roelfsema P, Changeux J-P *et al.* Conscious processing and the global neuronal workspace hypothesis. *Neuron* 2020;**105**:776–98.

Massimini M, Boly M, Casali A *et al.* A perturbational approach for evaluating the brain's capacity for consciousness. *Prog Brain Res* 2009;**177**:201–14.

Merikle PM, Smilek D, Eastwood JD. Perception without awareness: perspectives from cognitive psychology. *Cognition* 2001;**79**:115–34.

Moore T, Zirnsak M. Neural mechanisms of selective visual attention. *Ann Rev Psychol* 2017;**68**:47–72.

Rosanova M, Gosseries O, Casarotto S *et al.* Recovery of cortical effective connectivity and recovery of consciousness in vegetative patients. *Brain* 2012;**135**:1308–20.

Rosenthal D. *Consciousness and Mind*. UK: Oxford University Press, 2005.

Schurger A, Cowey A, Tallon-Baudry C. Induced gamma-band oscillations correlate with awareness in hemianopic patient GY. *Neuropsychologia* 2006;**44**:1796–803.

Searle JR. *The Mystery of Consciousness*. New York: New York Review of Books, 1997.

Sergent C, Baillet S, Dehaene S. Timing of the brain events underlying access to consciousness during the attentional blink. *Nat Neuro* 2005;**8**:1391–400.

Simons DJ, Chabris CF. Gorillas in our midst: sustained inattentional blindness for dynamic events. *Perception* 1999;**28**:1059–74.

Stoerig P, Cowey A. Blindsight in man and monkey. *Brain* 1997;**120**:535–59.

Tononi G, Boly M, Massimini M *et al.* Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci* 2016;**17**:450–61.

Tononi G, Edelman GM. Consciousness and complexity. *Science* 1998;**282**:1846–51.

Tsushima Y, Sasaki Y, Watanabe T. Greater disruption due to failure of inhibitory control on an ambiguous distractor. *Science* 2006;**314**:1786–8.

Webb TW, Graziano MSA. The attention schema theory: a mechanistic account of subjective awareness. *Front Psychol* 2015;**6**:500.

Webb TW, Kean HH, Graziano MSA. Effects of awareness on the control of attention. *J Cogn Neurosci* 2016;**28**:842–51.

Whyte CJ, Smith R. The predictive global neuronal workspace: a formal active inference model of visual consciousness. *Prog Neurobiol* 2021;**199**:101918.

Wilterson AI, Graziano MSA. The attention schema theory in a neural network agent: controlling visuospatial attention using a descriptive model of attention. *Proc Natl Acad Sci U S A* 2021;**118**:e2102421118.

Wilterson AI, Kemper CM, Kim N *et al*. Attention control and the attention schema theory of consciousness. *Prog Neurobiol* 2020;**195**:101844.

Woodman GF, Luck SJ. Dissociations among attention, perception, and awareness during object-substitution masking. *Psychol Sci* 2003;**14**:605–11.

Woodward J, Ross L. Scientific explanation. In: Zalta EN (ed.), *Stanford Encyclopedia of Philosophy*. Summer 2021 edn. Stanford, CA: Metaphysics Research Lab, Stanford University, 2021.