13

THE ATTENTION SCHEMA THEORY OF CONSCIOUSNESS

Michael S. Graziano

Over the past several years, my colleagues and I outlined a novel approach to understanding the brain basis of consciousness. That approach was eventually called the Attention Schema Theory (AST) (Graziano 2010; Graziano and Kastner 2011; Graziano 2013; Graziano 2014; Kelly at al. 2014; Webb and Graziano 2015; Webb, Kean, and Graziano 2016; Webb et al. 2016). The core concept is extremely simple. The brain not only uses the process of attention to focus its resources onto select signals, but it also constructs a description, or representation, of attention. The brain is a model builder – it builds models of items in the world that are useful to monitor and predict. Attention, being an important aspect of the self, is modeled by an attention schema.

The hypothesized attention schema is similar to the body schema. The brain constructs a rough internal model or simulation of the body, useful for monitoring, predicting, and controlling movement (Graziano and Botvinick 2002; Holmes and Spence 2004; Macaluso and Maravita 2010; Wolpert et al. 1995). Just so, the brain constructs a rough model of the process of attention – what it does, what its most basic properties are, and what its consequences are.

In the theory, the internal model of attention is a high-level, general description of attention. It lacks a description of the physical nuts and bolts that undergird attention, such as synapses, neurons, and competing electrochemical signals. The model incompletely and incorrectly describes the act of attending to X as, instead, an ethereal, subjective awareness of X. Because of the information in that internal model, and because the brain knows only the information available to it, people describe themselves as possessing awareness and have no way of knowing that this description is not literally accurate.

Although AST may seem quite different from other theories of consciousness, it is not necessarily a rival. Instead, I suggest it is compatible with many of the common, existing theories, and can add a crucial piece that fills a logical gap. Most theories of consciousness suffer from what might be called the metaphysical gap. The typical theory offers a physical mechanism, and then makes the assertion, "and then subjective awareness happens." The bridge between a physical mechanism and a metaphysical experience is left unexplained. In contrast, AST has no metaphysical gap, because it contains nothing metaphysical. Instead its explanation arrives at the step, "And then the machine claims that it has subjective awareness; and its internal computations consistently and incorrectly loop to the conclusion that this self-description is literally accurate." Explaining how a machine computes information is a matter of engineering, not a matter of

metaphysics. Even if many of the steps have not yet been filled in, none present a fundamental, scientifically unapproachable mystery.

In this chapter, I summarize AST and then discuss some of the ways it might make contact with three specific approaches to consciousness: higher-order thought, social theories of consciousness, and integrated information. This chapter does not review the specific experimental evidence in support of AST, described in other places (Kelly at al. 2014; Webb and Graziano 2015; Webb, Kean, and Graziano 2016; Webb et al. 2016). Instead it summarizes the concepts underlying the theory.

1 Awareness

AST posits a specific kind of relationship between awareness and attention. Explaining the theory can be difficult, however, because those two key terms have an inconvenient diversity of definitions and connotations. The next few sections, therefore, focus on explaining what I mean by "awareness" and "attention."

When people say, "I am aware of X," whatever X may be – a touch on the skin, an emotion, a thought – they typically mean that X is an item within subjective experience, or in mind, at that moment in time. This is the sense in which I use the term in this chapter. To be aware is to have a subjective experience.

The term is also sometimes used in another sense: If someone asks, "Are you aware that paper is made from trees?" you might say, "Of course I am." You are aware in the sense that the information was available in your memory. But by the definition of the word that I use in this chapter, you were not aware of it while it was latent in your memory. You became aware of it – had a subjective experience of thinking it – when you were reminded of the fact, and then you stopped being aware of it again when it slipped back out of your present thought.

A third, less common use of the word, "objective awareness," is found in the scientific literature (Lau 2008). The essential concept is that if the information gets into a person's brain and is processed in a manner that is objectively measurable in the person's behavior, then the person is "objectively aware" of the information. In this sense, one could say, "My microwave is aware that it must stop cooking in thirty seconds." Objective awareness has no connotation of an internal, subjective experience.

In this chapter, when I use the term awareness, I do not mean objective awareness. I also do not mean something that is latent in memory. I am referring to the moment-by-moment, subjective experience. Some scholars refer to this property as "consciousness." Some, in an abundance of zeal, call it "conscious awareness." In this chapter, for simplicity, I will use the term "awareness." One can have awareness of a great range of items, from sensory events to abstract thoughts.

The purpose of AST is to explain how the human brain claims to have so quirky and seemingly magical a property as an awareness of some of its information content. This problem has sometimes been called the "hard problem" of consciousness (Chalmers 1996).

2 Attention

The term "attention" has even more meanings and interpretations than "awareness." Here, I will not be able to give a single definition, but will describe the general class of phenomenon that is relevant to AST.

First, I will clarify what I do *not* mean by attention. A typical colloquial use of the term conflates it with awareness. In that colloquial use, awareness is a graded property – you are

more vividly aware of some items than others – and the items of which you are most aware at any moment are the items within your attention. This meaning is close to William James' now famous definition of attention (James 1890): "It is the taking possession of the mind, in clear and vivid form, of one out of what seems several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence." In this intuitive approach, attention is part of subjective experience. It is a subset of the conscious mind. If the content of awareness is the food spread at a banquet, attention refers specifically to the food on the plate directly in front of you. However, that is not what I mean by attention.

In this chapter, I use the term "attention" to refer to a mechanistic process in the brain. It can be defined independently of any subjective experience, awareness, or mind. Attention is the process by which some signals in the brain are boosted and therefore processed more deeply, often at the expense of other competing signals that are partially suppressed. Attention is a data-handling process. It can be measured in a great variety of ways, including through faster reaction times and greater accuracy in understanding, remembering, and responding to an attended item.

Many different kinds of attention have been described by psychologists (for review, see Nobre and Kastner 2014). Psychologists have distinguished between overt attention (turning the head and eyes toward a stimulus) and covert attention (focusing one's processing on a stimulus without looking directly at it). Psychologists have also distinguished between bottom-up, stimulus-driven attention (such as to a flashing light) and top-down, internally driven attention (such as looking for a friend in a crowd). Other categorizations include spatial attention (enhancing the sensory signals from a particular location in space) and object attention (enhancing the processing of one object over another, even if the two are superimposed on each other at the same spatial location). One can direct visual attention, auditory attention, tactile attention, and even multisensory attention. It has been pointed out that people can focus attention on specific abstract thoughts, beliefs, memories, or emotions, events that are generated in the brain and that are not directly stimulus-linked (Chun et al. 2011).

One of the most influential perspectives on attention, is a neuroscientific account called the biased competition model (Desimone and Duncan 1995; Beck and Kastner 2009). In that account, the relevant signals – whether visual, auditory, or anything else – are in competition with each other. The competition is driven ultimately by synaptic inhibition among interconnected neurons. Because of this inhibition, when many signals are in competition, one will tend to rise in strength and suppress the others. That competition is unstable – shifting from one winner to another, from one moment to the next, depending on a variety of influences that may tip or bias the competition. The biasing influences include bottom-up, stimulus-driven factors (such as the brightness of a stimulus) and top-down, internally generated factors (such as a choice to search a particular location). The biased competition model provides a neuronal mechanism that explains how some signals become enhanced at the expense of others.

Attention is clearly a complex, multifaceted process. It is probably best described as many different processes occurring at many levels in the brain, applied to many information domains. Yet there is a common thread among these many types of attention. Throughout this chapter, when I use the term attention, I am referring to the selective enhancement of some signals in the brain over other competing signals, such that the winning signals are more deeply processed and have a greater impact on action, memory, and cognition.

3 Comparing Awareness to Attention

The relationship between awareness and attention has been discussed many times before (e.g. Koch and Tsuchiya 2007; Lamme 2004). A variety of theories of consciousness emphasize that

relationship (e.g. Prinz 2012). In AST, one specific kind of relationship is hypothesized. To better explain that proposed relationship, in this section I list eight similarities and two differences between attention and awareness. The subsequent section will discuss why that list of similarities and differences suggests a specific kind of relationship between attention and awareness.

Similarity 1: Both involve a target. You attend *to* something. You are aware *of* something.

Similarity 2: Both involve a source. Attention is a data-handling operation performed by the processing elements in a brain. Awareness implies an "I," an agent who is aware.

Similarity 3: Both are selective. Only some of the available information is attended at any one time, or enters awareness at any one time.

Similarity 4: Both have an uneven, graded distribution, typically with a single focus. While attending mostly to A, the brain can spare some attention for B, C, and D. One can be most intently aware of A and a little aware of B, C, and D.

Similarity 5: Both imply deep processing. Attention is when an information processor devotes computing resources to a selected signal and thereby arrives at a deeper or more detailed encoding of it. Awareness implies an intelligence seizing on, being occupied by, knowing or experiencing something.

Similarity 6: Both imply an effect on behavior and memory. When the brain attends to something, the enhanced neural signals have a greater impact on behavioral output and memory. When the brain does not attend to something, the neural representation is weak and has relatively little impact on behavior or memory. Likewise, when you are aware of something, by implication you can choose to act on it and are able to remember it. When you are unaware of something, by implication, you probably fail to react to it or remember it.

Similarity 7: Both operate on similar domains of information. Although most studies of attention focus on vision, it is certainly not limited to vision. The same signal enhancement can be applied to signals arising in any of the five senses– to a thought, to an emotion, to a recalled memory, or to a plan to make a movement, for example. Just so, one can be aware of the same range of items. Generally, if you can in principle direct attention to it, then you can in principle be aware of it, and *vice versa*.

Similarity 8: Not only *can* attention and awareness apply to the same item, they almost always do. Here the relationship is complex. It is now well established that attention and awareness can be dissociated (Hsieh et al. 2011; Jiang et al. 2006; Kentridge et al. 2008; Koch and Tsuchiya 2007; Lambert 1988; Lambert et al. 1999; Lamme 2004; McCormick 1997; Norman et al. 2013; Tsushima et al. 2006; Webb, Kean, and Graziano 2016). A great many experiments have shown that people can pay attention to a visual stimulus, in the sense of processing it deeply, and yet at the same time have no subjective experience of the stimulus. They insist they cannot see it. This dissociation shows that attention and awareness are not the same. Awareness is not merely "what it feels like" to pay attention. Arguably, this point could be labeled "Difference 1" rather than "Similarity 8." However, the dissociation between attention and awareness should not be exaggerated. It is surprisingly difficult to separate the two. The dissociation seems to require either cases of brain damage, or visual stimuli that are extremely dim or masked by other stimuli, such that they are near the threshold of detection. Only in degraded conditions is it possible to reliably separate attention from awareness. Under most conditions, awareness and

attention share the same target. What you attend to, you are usually aware of. This almost-but-not-quite registration between awareness and attention plays a prominent role in AST.

Awareness and attention are so similar that it is tempting to conclude that they are simply different ways of measuring the same thing, and that the occasional misalignment is caused by measurement noise. However, I find at least two crucial difference that are important in AST.

Difference 1: We know scientifically that attention is a process that includes many specific, physical details. Neurons, synpases, electrochemical signals, ions and ion channels in cell membranes, a dance of inhibitory and excitatory interactions, all participate in the selective enhancement of some signals over others. But awareness is different: we describe it as a thing that has no physical attributes. The awareness stuff itself isn't the neurons, the chemicals, or the signals–although we may think that awareness arises from those physical underpinnings. Awareness itself is not a physical thing. You cannot push on it and measure a reaction force. It is a substanceless, subjective feeling. In this sense, awareness, as most people conceptualize it, is metaphysical. Indeed, the gap between physical mechanism and metaphysical experience is exactly why awareness has been so hard to explain.

Difference 2: Attention is something the brain demonstrably *does* whereas awareness is something the brain *says that it has*. Unless you are a neuroscientist with a specific intellectual knowledge, you are never going to report the state of your actual, mechanistic attention. Nobody ever says, "Hey, you know what just happened? My visual neurons were processing both A and B, and a competition ensued in which lateral inhibition, combined with a biasing boost to stimulus A, caused..." People do not report directly on their mechanistic attention. They report on the state of their awareness. Even when people say, "I'm paying attention to that apple," they are typically using the word "attention" in a colloquial sense, not the mechanistic sense as I defined it above. In the colloquial sense of the word, people typically mean, "My conscious mind is focusing on that apple; it is uppermost in my awareness." Again, they are reporting on the state of their awareness, not on their mechanistic process of attention.

In summary, awareness and attention match point-for-point in many respects. They seem to have similar basic properties and dynamics. They are also tightly coupled in most circumstances, becoming dissociated from each other only at the threshold of sensory performance. But attention is a physically real, objectively measurable event in the brain, complete with mechanistic details, whereas awareness is knowledge that can be reported, and we report it as lacking physical substance or mechanistic details.

This pattern of similarities and differences suggests a possible relationship between attention and awareness: awareness is the brain's incomplete, detail-poor description of its own process of attention. To better grasp what I mean by this distinction between attention (a physically real item) and awareness (a useful if incomplete description of attention), consider the following examples. A gorilla is different from a written report about gorillas. The book may contain a lot of information, but is probably incomplete, perhaps even inaccurate in some details. An apple is different from the image of an apple projected onto your retina. An actual clay pipe is not the same as Magritte's famous oil painting of a pipe that he captioned, "This is not a pipe." The next section describes this hypothesized relationship in greater detail.

4 Analogy to the Body Schema

To better explain the possible relationship between attention and awareness, I will use the analogy of the body and the body schema (Graziano and Botvinick 2002; Holmes and Spence 2004; Macaluso and Maravita 2010; Wolpert et al. 1995). Imagine you close your eyes and tell me about your right arm – not what you know intellectually about arms in general, but what you can tell about your particular arm, at this particular moment, by introspection. What state is it in? How is it positioned? How is it moving? What is its size and shape? What is the structure inside? How many muscles do you have inside your arm and how are they attached to the bones? Can you describe the proteins that are cross-linking at this moment to stiffen the muscles?

You can answer some of those questions, but not all. General information about the shape and configuration of your arm is easy to get at, but you can't report the mechanistic details about your muscles and proteins. You may even report incorrect information about the exact position of your arm. The reason for your partial, approximate description, is that you are not reporting on your actual arm. Your cognitive machinery has access to an internal model, a body schema, that provides incomplete, simplified information about the arm. You can report some of the information in that arm schema. Your cognition has access to a repository of information, an arm model, and the arm model is simplified and imperfect.

My point here is to emphasize the specific, quirky relationship between the actual arm and the arm schema. In AST, the relationship between attention and awareness is similar. Attention is an actual physical process in the brain, and awareness is the brain's constantly updated model of attention.

Suppose you tell me that you are aware of item X – let's say an apple placed in front of you. In AST, you make that claim of awareness because you have two closely related internal models. First, you have an internal model of the apple, which allows you to report the properties of the apple. You can tell me that it's round, it's red, it's at a specific location, and so on. But that by itself is not enough for awareness. Second, you have an internal model of attention, which allows you to report that you have a specific kind of mental relationship to the apple. When you describe your awareness of the apple – the mental possession, the focus, the non-physical subjective experience – according to AST, that information comes from your attention schema, a rough, detail-poor description of your process of attention.

5 Why An Attention Schema Might Cause a Brain to Insist That It Has Subjective Awareness – and Insist That It Isn't Just Insisting

Suppose you play me for a fool and tell me that you are literally an iguana. In order to make that claim, you must have access to that information. Something in your brain has constructed the information, "I am an iguana." Yet that information has a larger context. It is linked to a vast net of information to which you have cognitive access. That net of information includes much that you are not verbalizing to me, including the information, "I'm not really an iguana," "I made that up just to mess with him," "I'm a person," and so on. Moreover, that net of information is layered. Some of it is at a cognitive level, consisting of abstract propositions. Some of it is at a linguistic level. Much is at a deeper, sensory or perceptual level. You have a body schema that informs you of your personhood. Your visual system contains sensory information that also confirms your real identity. You have specific memories of your human past.

But, suppose I am cruelly able to manipulate the information in your brain, and I alter that vast set of information to render it consistent with the proposition that you are an iguana. Your body schema is aligned to the proposition. So is the sensory information in your visual system,

and the information that makes up your memory and self-knowledge. I remove the specific information that says, "I made that up just to mess with him." I switch the information that says, "I am certain this is not true," to its opposite, "I'm certain it's true." Now how can you know that you are not an iguana? Your brain is captive to the information it contains. Tautologically, it knows only what it knows. You would no longer think of your iguana identity as hypothetical, or as mere information at an intellectual level. You would consider it a ground truth.

Now we can explain the widespread human conviction that we have an inner, subjective experience. In AST, the attention schema is a set of information that describes attention. It does not describe the object you are attending to – that would be a different schema. Instead it describes the act of attention itself. Higher cognition has a partial access to that set of information, and can verbally report some of its contents.

Suppose you are looking at an apple and I ask you, "Tell me about your awareness of the apple – not the properties of the apple, but the properties of the awareness itself. What is this awareness you have?" Your cognitive machinery, gaining access to the attention schema, reports on some of the information within it. You answer, "My mind has taken hold of the apple. That mental possession empowers me to know about the apple, to remember it for later, to act on it."

"Fair enough," I say, "but tell me about the physical properties of this awareness stuff." Now you're stuck. That internal model of attention lacks a description of any of the physical details of neurons, synapses, or competing signals. Your cognition, reporting on the information available to it, says, "The awareness itself has no physically describable attributes. It just is. It's a non-physical essence located inside me. In that sense, it's metaphysical. It's the inner, mental, experiential side of me."

The machine, based on an incomplete model of attention, claims to have a subjective experience.

I could push you further. I could say, "But you're just a machine accessing internal models. Of course, you're going to say all that, because that's the information contained in those internal models." Your cognition, searching the available internal models, finds no information that matches that description. Nothing in your internal models say, "This is all just information in a set of internal models." Instead, you reply, "What internal models? What information? What computation? No, simply, there's a me, there's an apple, and I have a subjective awareness of the apple. It's a ground truth. It simply exists."

This is a brain stuck in a loop, captive to the information available to it.

AST does not explain how the brain generates a subjective inner feeling. It explains how a brain *claims* to have a subjective inner feeling. In this theory, there is no awareness essence that arises from the functioning of neurons. Instead, in AST, the brain contains attention. Attention is a mechanistic, data-handling process. The brain also constructs an incomplete and somewhat inaccurate internal model, or description, of attention. On the basis of that internal model, the brain insists that it has subjective awareness – and insists that it is not just insisting. That general approach, in which awareness does not exist as such, and our claim to have awareness can be cast in terms of mechanistic information processing, is similar to the general approach proposed by Dennett (1991).

In AST, awareness is not merely an intellectual construct. It is an automatic, continuous, fundamental construct about the self, to which cognition and language have partial access.

6 Three Ways in Which the Theory Remains Incomplete

AST is underspecified in at least three major ways, briefly summarized in this section.

First, if the brain contains an attention schema, which of the many kinds of attention does it model? There are many overlapping mechanisms of attention, as noted in an earlier section

of this chapter. These mechanisms operate at many levels, from the lowest sensory processing levels to the highest levels of cognition. If the brain has an attention schema, does it model only one type of attention? Many types? Are there many attention schemas, each modeling a different mix of attention mechanisms? In its current provisional form (Graziano 2013; Webb and Graziano, 2015), the theory posits that a single attention schema models an amalgam of all levels of attention. In that view, the reality of attention is a complex and layered process, but the attention schema depicts it in a simplified manner as a single amorphous thing – an awareness.

A second way in which AST is not yet fully specified concerns the information content of the attention schema. It is extremely difficult to specify the details of an information set constructed in the brain. In the case of the body schema, for example, after a hundred years of study, researchers have only a vague understanding of the information contained within it (Graziano and Botvinick 2002; Holmes and Spence 2004; Macaluso and Maravita 2010; Wolpert et al. 1995). It contains information about the general shape and structure of the body, as well as information about the dynamics of body-movement. In the case of the attention schema, if the brain is to construct an internal model of attention, what information would be useful to include? Perhaps basic information about the properties of attention – it has an object (the target of attention); it is generated by a subject (the agent who is attending); it is selective; it is graded; it implies a deep processing of the attended item; and it has specific, predictable consequences on behavior and memory. Perhaps the attention schema also includes some dynamic information about how attention tends to move from point to point and how it is affected by different circumstances. The fact is, at this point, the theory provides very little indication of the contents of the attention schema. Only future work will be able to fill in those details.

The third way in which AST is underspecified concerns the functions of an attention schema. Why would such a thing evolve? A range of adaptive functions are possible. For example, an attention schema could in principle be used for controlling one's own attention (Webb and Graziano, 2015; Webb, Kean, and Graziano 2016). By analogy, the brain constructs the internal model of the arm to help control arm movements (e.g. Haith and Krakauer 2013; Scheidt et al. 2005; Wolpert et al. 1995). It is a basic principle of control engineering (Camacho and Bordons Alba 2004).

A possible additional function of an attention schema, is to model the attentional states of other people (Kelly et al. 2014; Pesquita et al. 2016). The more a person attends to X, the more likely that person is to react to X. Modeling attention is therefore a good way to predict behavior. By attributing awareness to yourself and to other people, you are in effect modeling the attentional states of interacting social agents. You gain some ability to predict everyone's behavior including your own. In this way, an attention schema could be fundamental to social cognition.

7 Higher-Order Thought

The higher-order thought theory, elaborated by Rosenthal, is currently one of the most influential theories of consciousness (Lau and Rosenthal 2011; Rosenthal 2005; Gennaro 1996, 2012). I will briefly summarize some of its main points and note its possible connection to AST.

Consider how one becomes aware of a visual stimulus such as an apple. In the higher-order thought theory, the visual system constructs a sensory representation of the apple. Higher-order systems in the brain receive that information and re-represent the apple. That higher-order re-representation contains the extra information that causes us to report not only the presence of the apple, but also a subjective experience.

The higher-order thought theory is a close cousin of AST because of its focus on representation and information. The theory, however, focuses on the representation of the item (such as the apple in the example above) that is within awareness, in contrast to AST which focuses on the representation of the process of attention.

Higher-order thought theory is surprisingly compatible with AST. In the combination theory, the brain constructs a representation of the apple. It also constructs a representation of attention – the attention schema. A higher-order re-representation combines the two. That higher-order representation describes an apple to which one's subjective awareness is attached. Given that higher-order representation, the system can make two claims. First, it can report the properties of the apple. Second, it can report a subjective awareness associated with the apple. By adding an attention schema to the mix, we add the necessary information for the machine to report awareness – otherwise, the machine would have no basis for even knowing what awareness is or concluding that it has any. In this perspective, AST is not a rival to the higher-order thought theory. Instead, the two approaches synergize and gain from each other.

8 Social Attribution of Awareness

Recently, Prinz (2017) outlined a view of consciousness termed import theory. In that perspective, humans first develop the ability to model the mind states of others and then turn that ability inward, attributing similar mind states to themselves. This explanation of conscious mind states, invoking social cognition, has been proposed before many times in different forms, including in the earliest descriptions of AST (Graziano 2013), but Prinz presents the view in a particularly clear and compelling manner.

One of the strengths of the import theory, is that it covers a broad range of mind states, all of which compose what most people colloquially think of as consciousness. You can attribute emotions, thoughts, goals, desires, beliefs, and intentions to other people. Just so, you can attribute the same range of mind states to yourself. The theory therefore addresses a rich world of consciousness that is often ignored in discussions of sensory awareness.

However, the theory has the same metaphysical gap as so many other theories contain. It addresses the content of awareness, but it does not address how we get to be aware of it. You may attribute an emotional state to another person, and you may attribute the same emotional state to yourself. But why do you claim to have a subjective *experience* of that emotion? It is not enough for the brain, computer-like, to build the construct, "I am happy." Humans also report a subjective experience of the happiness, just as they report a subjective experience of many other items. Import theory, by itself, does not explain the subjective experience. This point is not meant as a criticism of the theory. It is a valuable theory – but the specific question of awareness may lie outside its domain.

AST may be able to fill that gap. In AST, when we attribute awareness to another person, we are modeling that person's state of attention. When we attribute awareness to ourselves, we are modeling our own state of attention. By adding an attention schema to the system, we add information that allows the brain to know what awareness is in the first place and to claim that it has some, or that someone else has some. Note that, strictly speaking, AST does not explain how people *have* subjective awareness. It explains how people *insist* that they have it and insist that it's real and that they're not just insisting.

I do not mean to take a strong stand here on import theory, for or against. It is possible that people develop the ability to model the mind states of others first and then import that to the

self. It is also possible that people develop the capacity of self-modeling first and then export it outward to others. Maybe both are true. Only more data will be able to untangle those possibilities. My point here is that, whichever perspective one prefers, AST makes a useful addition.

A skeptical colleague might wonder, "Why focus on attention, when the brain contains so many different processes? Decisions, emotions, moods, beliefs – all of these are a part of consciousness. Yes, surely the brain constructs a model of attention, but doesn't it also construct models of all its other internal processes?" Indeed, the brain probably does construct models of many internal processes, and all of those models are worthy of scientific study. The reason AST highlights attention is that an attention schema answers one crucial, focused question that was thought to be unanswerable. It explains how people claim to have a subjective experience of anything at all. Because of the narrow specificity of AST, it can be added as a useful component to a great range of other theories.

9 Networked Information

Many theories and speculations about awareness share an emphasis on the widespread networking or linking of information around the brain. Two prominent examples are the Integrated Information Theory (Tononi 2008) and the Global Workspace Theory (Baars 1988; Dehaene 2014).

The essence of the Integrated Information Theory is that if information is integrated to a sufficient extent, which may be mathematically definable, then subjective awareness of that information is present (Tononi 2008). Awareness is what integrated information feels like. The Global Workspace Theory has at least some conceptual similarities (Baars 1988; Dehaene 2014). You become subjectively aware of a visual stimulus, such as an apple, because the representation of the apple in the visual system is globally broadcasted and accessible to many systems around the brain. Again, the widespread sharing of information around the brain results in awareness. Many other researchers have also noted the possible relationship between awareness and the binding, integration, or sharing of information around the brain (e.g. Crick and Koch 1990; Damasio 1990; Engel and Singer 2001; Lamme 2006).

Of all the common theories of consciousness in the cognitive psychology literature, this class of theory most obviously suffers from a metaphysical gap. To explain an awareness of item X, these theories focus on the information about X and how that information is networked or integrated. The awareness is treated as an adjunct, or a symptom, or a product, of the information about X. But once you have information that is integrated, or that is globally broadcasted, or that is linked or bound across different domains, why would it take the next step and enter a state of subjective awareness? Why is it not just a pile of integrated information without the subjective experience? What is the actual awareness stuff and how does it emerge from that state of integration?

Another way to put the question is this: Suppose you have a computing machine that contains information about an apple. Suppose that information is highly-integrated – color, shape, size, texture, smell, taste, identity, all cross-associated and integrated in a massive brain-wide representation. I can understand how a machine like that might be able to report the properties of the apple, but why would I expect the machine to add to its report, "And by the way, I have a subjective, internal *experience* of those apple properties"? What gave the machine the informational basis to report a subjective experience?

The metaphysical gap has stood in the way of these theories that depend on networked information. And yet the conundrum has a simple solution. Add AST to the integrated information account, and you have a working theory of awareness. If part of the information that is integrated globally around the brain consists of information about awareness, about what awareness is, what its properties are, about how you yourself are aware and what in specific you are aware

of – if the machine contains an attention schema – then it is equipped to talk about awareness in all its subtle properties and to make the claim that it has those properties. If the machine lacks information about awareness, then logically it cannot claim to have any.

Note that not only is AST a useful addition to the integrated information perspective, but the relationship works both ways. AST depends on integrated information. It does not work as a theory without the widespread networking of information around the brain. In AST, to be aware of an apple, it is not enough to construct an attention schema. The attention schema models the properties of attention itself. The brain must also construct an internal model of the apple and an internal model of the self as a specific agent. All three must be integrated across widely divergent brain areas, building a larger internal model. That overarching, integrated internal model contains the information: there is a you as an agent with a set of specific properties, there is an apple with its own set of specific properties, and at this moment the you-as-agent has a subjective awareness of the apple and its properties. Only with that highly networked information is the brain equipped to claim, "I am aware of the apple." Without the widespread integration of information around the brain, that overarching internal model is impossible, and we would not claim to possess awareness. Thus, even though AST and the integrated information approach rest on fundamentally different philosophical perspectives, they have a peculiarly close, symbiotic relationship.

10 The Allure of Introspection

Before Newton's publication on light (1671), the physical nature of color was not understood. White light was assumed to be pure and colored light to be contaminated. One could say the hard problem of color was this: how does white light become scrubbed clean of contaminants? That hard problem, alas, had no answer because it was based on a physically incoherent model of color and light. The model was not merely a mistaken scientific theory. It was the result of millions of years of evolution working on the primate visual system, shaping an efficient and simplified internal model of reflectance spectrum. Finally, after Newton's insights, it became possible to understand two crucial items. First, white light is actually a mixture of all colors. Second, the model we all automatically construct in our visual systems is simplified and in some respects wrong.

The same issues, I suggest, apply to the study of awareness. Our cognitive machinery gains partial access to deeper internal models, including an attention schema. On the basis of that information, people assert with absolute confidence that they have physically incoherent, magicalist properties. Gradually, as science has made progress over hundreds of years, some of the more obviously irrational assertions have fallen away. Most scientists accept there is no such thing as a ghost. A mysterious energy does not emanate from the eyes to affect other objects and people. Most neuroscientists reject the dualist notion of mind and brain, the notion most famously associated with Descartes (1641), in which the machine of the brain is directed by the metaphysical substance of the mind.

Some of the assertions of magic, however, remain with us in subtle ways. Almost all theories of consciousness rest on a fundamental assumption: we have an inner subjective experience. The experience is not itself a physical substance. It cannot be weighed, poked, or directly measured. You cannot push on it and measure a reaction force. Instead it is a non-physical, side-product – the "what-it-feels-like" when certain processes occur in the brain. The challenge is to explain how the functioning of the brain results in that private feeling.

This perspective has framed the entire field of consciousness studies from the beginning. Yet, I argue it is as futile as the attempt to explain how white light becomes purified of contaminants. It is predicated on false assumptions. As long as we dedicate ourselves to explaining how the

brain produces subjective experience, a property we know about only by our cognition accessing our internal models, we will never find the answer. As soon as we step away from the incorrect assumptions, and realize that our evolutionarily built-in models are not literally accurate, we will see that the answer to the question of consciousness is already here.

The heart of AST is that the brain is a machine: it processes information. When we claim to have a subjective experience, and swear on it, and vociferously insist that it isn't just a claim or a conclusion – it's real, dammit – this output occurs because something in the brain computed that set of information. It is a self-description. The self-model is unlikely to be entirely accurate or even physically coherent. As in the case of color, the brain's models tend to be efficient, simplified, useful, and not very accurate on those dimensions where accuracy would serve no clear behavioral advantage. People do not have a magic internal feeling. We have information that causes us to insist that we have the magic. And explaining how a machine computes and handles information is well within the domain of science.

References

Baars, B. J. (1988) *A Cognitive Theory of Consciousness*, New York: Cambridge University Press.

Beck, D. M. and Kastner, S. (2009) "Top-down and bottom-up mechanisms in biasing competition in the human brain," *Vision Research* 49: 1154–1165.

Camacho, E. F. and Bordons Alba, C. (2004) *Model Predictive Control*, New York: Springer.

Chalmers, D. (1996) *The Conscious Mind*, New York: Oxford University Press.

Chun, M. M., Golomb, J. D. and Turk-Browne, N. B. (2011). "A taxonomy of external and internal attention," *Annual Review of Psychology* 62: 73–101.

Crick, F. and Koch, C. (1990) "Toward a neurobiological theory of consciousness," *Seminars in the Neurosciences* 2: 263–275.

Damasio, A. R. (1990) "Synchronous activation in multiple cortical regions: a mechanism for recall," *Seminars in the Neurosciences* 2: 287-296.

Dehaene, S. (2014) *Consciousness and the Brain*, New York: Viking.

Dennett, D. C. (1991) *Consciousness Explained*, Boston: Little, Brown, and Co.

Descartes, R. (1641) "Meditations on first philosophy," in J. Cottingham, R. Stoothoff, and D. Murdock (trans.) *The Philosophical Writings of Rene Descartes*, Cambridge: Cambridge University Press.

Desimone, R. and Duncan, J. (1995) "Neural mechanisms of selective visual attention," *Annual Review of Neuroscience* 18: 193–222.

Engel, A. K. and Singer, W. (2001) "Temporal binding and the neural correlates of sensory awareness," *Trends in Cognitive Sciences* 5: 16–25.

Gennaro, R. (1996) *Consciousness and Self Consciousness: A Defense of the Higher Order Thought Theory of Consciousness*, Philadelphia, PA: John Benjamins Publishing.

Gennaro, R. (2012) *The Consciousness Paradox: Consciousness, Concepts, and Higher-Order Thoughts*, Cambridge, MA: The MIT Press.

Graziano, M. S. A. (2010) *God, Soul, Mind, Brain: A Neuroscientists Reflections on the Spirit World*, Fredonia: Leapfrog Press.

Graziano, M. S. A. (2013) *Consciousness and the Social Brain*, New York: Oxford University Press.

Graziano, M. S. A. (2014) "Speculations on the evolution of awareness," *Journal of Cognitive Neuroscience* 26: 1300–1304.

Graziano, M. S. A. and Botvinick, M. M. (2002) "How the brain represents the body: insights from neurophysiology and psychology," in W. Prinz and B. Hommel. (eds.) *Common Mechanisms in Perception and Action: Attention and Performance XIX*, Oxford: Oxford University Press.

Graziano, M. S. A. and Kastner, S. (2011) "Human consciousness and its relationship to social neuroscience: A novel hypothesis," *Cognitive Neuroscience* 2: 98–113.

Haith A. M. and Krakauer, J.W. (2013) "Model-Based and Model-Free Mechanisms of Human Motor Learning," in M. Richardson, M. Riley, and K. Shockley (eds.) *Progress in Motor Control: Advances in Experimental Medicine and Biology, Vol* 782, New York: Springer.

Holmes, N. and Spence, C. (2004) "The body schema and the multisensory representation(s) of personal space," *Cognitive Processing* 5: 94–105.

Hsieh, P., Colas, J. T. and Kanwisher, N. (2011) "Unconscious pop-out: Attentional capture by unseen feature singletons only when top-down attention is available," *Psychological Science* 22: 1220–1226.
James, W. (1890) *Principles of Psychology*, New York: Henry Holt and Company
Jiang, Y., Costello, P., Fang, F., Huang, M. and He, S. (2006) "A gender- and sexual orientation-dependent spatial attentional effect of invisible images," *Proceedings of the National Academy of Sciences U. S. A.* 103: 17048–17052.
Kelly, Y. T., Webb, T. W., Meier, J. D., Arcaro, M. J. and Graziano, M. S. A. (2014) "Attributing awareness to oneself and to others," *Proceedings of the National Academy of Sciences U. S. A.* 111: 5012–5017.
Kentridge, R. W., Nijboer, T. C. and Heywood, C. A. (2008) "Attended but unseen: visual attention is not sufficient for visual awareness," *Neuropsychologia* 46: 864–869.
Koch, C. and Tsuchiya, N. (2007) "Attention and consciousness: two distinct brain processes," *Trends in Cognitive Sciences* 11: 16–22.
Lambert, A. J., Beard, C. T. and Thompson, R. J. (1988) "Selective attention, visual laterality, awareness and perceiving the meaning of parafoveally presented words," *Quarterly Journal of Experimental Psychology: Human Experimental Psychology* 40A: 615–652.
Lambert, A., Naikar, N., McLachlan, K. and Aitken, V. (1999) "A new component of visual orienting: Implicit effects of peripheral information and subthreshold cues on covert attention," *Journal of Experimental Psychology: Human Perception and Performance* 25: 321–340.
Lamme, V. A. (2004) "Separate neural definitions of visual consciousness and visual attention; a case for phenomenal awareness," *Neural Networks* 17: 861–872.
Lamme, V. A. (2006) "Towards a true neural stance on consciousness," *Trends in Cognitive Sciences* 10: 494–501.
Lau, H. (2008) "Are we studying consciousness yet," in L. Weiskrantz and M. Davies (eds.) *Frontiers of Consciousness: Chichele Lectures*, Oxford: Oxford University Press.
Lau, H. and Rosenthal, D. (2011) "Empirical support for higher-order theories of consciousness," *Trends in Cognitive Sciences* 15: 365–373.
Macaluso, E. and Maravita, A. (2010) "The representation of space near the body through touch and vision," *Neuropsychologia* 48: 782–795.
McCormick, P. A. (1997) "Orienting attention without awareness," *Journal of Experimental Psychology: Human Perception and Performance* 23: 168–180.
Newton, I. A. (1671) "Letter of Mr. Isaac Newton, Professor of the Mathematicks in the University of Cambridge; Containing His New Theory about Light and Colors: Sent by the Author to the Publisher from Cambridge, Febr. 6. 1671/72; In Order to be Communicated to the Royal Society," *Philosophical Transactions Royal Society* 6: 3075–3087.
Norbre, K. and Kaster, S. (2014) *The Oxford Handbook of Attention*, New York: Oxford Univeristy Press.
Norman, L. J., Heywood, C. A. and Kentridge, R. W. (2013) "Object-based attention without awareness," *Psychological Science* 24: 836–843.
Prinz, J. J. (2012) *The Conscious Brain*, New York: Oxford University Press.
Prinz, W. (2017) "Modeling Self on Others: An Import Theory of Subjectivity and Selfhood," in *Consciousness and Cognition* (in press).
Pesquita, A., Chapman, C. S. and Enns, J. T. (2016) "Humans are sensitive to attention control when predicting others' actions," *Proceedings of the National Academia of Science U. S. A.* 113: 8669–8674.
Rosenthal, D. (2005) *Consciousness and Mind*, New York: Oxford University Press.
Scheidt, R. A., Conditt, M. A., Secco, E. L. and Mussa-Ivaldi, F. A. (2005) "Interaction of visual and proprioceptive feedback during adaptation of human reaching movements," *Journal of Neurophysiology* 93: 3200–3213.
Tononi, G. (2008) "Consciousness as integrated information: a provisional manifesto," *Biological Bulletin* 215: 216–242.
Tsushima, Y., Sasaki, Y. and Watanabe, T. (2006) "Greater disruption due to failure of inhibitory control on an ambiguous distractor," *Science* 314: 1786–1788.
Webb, T. W., Kean, H. H. and Graziano, M. S. A. (2016) "Effects of awareness on the control of attention," *Journal of Cognitive Neuroscience* 28: 842–851.
Webb, T. W., and Graziano, M. S. A. (2015) "The attention schema theory: a mechanistic account of subjective awareness," *Frontiers in Psychology* 6, article 500, doi:10.3389/fpsyg.2015.00500.
Webb, T. W., Igelström, K., Schurger, A. and Graziano, M. S. A. (2016) "Cortical networks involved in visual awareness independently of visual attention," *Proceedings of the National Academy of Sciences U. S. A.* 113: 13923–13928.
Wolpert, D. M., Ghahramani, Z. and Jordan, M. I. (1995) "An internal model for sensorimotor integration," *Science* 269: 1880–1882.

Related Topics

Consciousness and Attention
The Intermediate Level Theory of Consciousness
Representational Theories of Consciousness
The Global Workspace Theory
The Information Integration Theory
The Neural Correlates of Consciousness