frontiers in Robotics and AI

HYPOTHESIS AND THEORY
published: 14 November 2017
doi: 10.3389/frobt.2017.00060

The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness

*Michael S. A. Graziano**

Department of Psychology and Neuroscience, Princeton University, Princeton, NJ, United States

The purpose of the attention schema theory is to explain how an information-processing device, the brain, arrives at the claim that it possesses a non-physical, subjective awareness and assigns a high degree of certainty to that extraordinary claim. The theory does not address how the brain might actually *possess* a non-physical essence. It is not a theory that deals in the non-physical. It is about the computations that cause a machine to make a claim and to assign a high degree of certainty to the claim. The theory is offered as a possible starting point for building artificial consciousness. Given current technology, it should be possible to build a machine that contains a rich internal model of what consciousness is, attributes that property of consciousness to itself and to the people it interacts with, and uses that attribution to make predictions about human behavior. Such a machine would "believe" it is conscious and act like it is conscious, in the same sense that the human machine believes and acts.

Keywords: attention, awareness, body schema, internal model, visual attention

Edited by:
Antonio Chella,
Università degli Studi di Palermo,
Italy

Reviewed by:
Ronald A. Rensink,
University of British Columbia,
Canada
Paul Syverson,
United States Naval Research
Laboratory, United States

***Correspondence:**
Michael S. A. Graziano
graziano@princeton.edu

Specialty section:
This article was submitted to
Humanoid Robotics,
a section of the journal
Frontiers in Robotics and AI

Received: *16 August 2017*
Accepted: *01 November 2017*
Published: *14 November 2017*

Citation:
Graziano MSA (2017) The Attention Schema Theory: A Foundation for Engineering Artificial Consciousness.
Front. Robot. AI 4:60.
doi: 10.3389/frobt.2017.00060

INTRODUCTION

This article is part of a special issue on consciousness in humanoid robots. The purpose of this article is to summarize the attention schema theory (AST) of consciousness for those in the engineering or artificial intelligence community who may not have encountered previous papers on the topic, which tended to be in psychology and neuroscience journals. The central claim of this article is that AST is mechanistic, demystifies consciousness and can potentially provide a foundation on which artificial consciousness could be engineered. The theory has been summarized in detail in other articles (e.g., Graziano and Kastner, 2011; Webb and Graziano, 2015) and has been described in depth in a book (Graziano, 2013). The goal here is to briefly introduce the theory to a potentially new audience and to emphasize its possible use for engineering artificial consciousness.

The AST was developed beginning in 2010, drawing on basic research in neuroscience, psychology, and especially on how the brain constructs models of the self (Graziano, 2010, 2013; Graziano and Kastner, 2011; Webb and Graziano, 2015). The main goal of this theory is to explain how the brain, a biological information processor, arrives at the claim that it possesses a non-physical, subjective awareness and assigns a high degree of certainty to that extraordinary claim. The theory does not address how the brain might actually *possess* a non-physical essence. It is not a theory that deals in the non-physical. It is about the computations that cause a machine to make a claim and to assign a high degree of certainty to the claim. The theory is in the realm of science and engineering.

Given a mechanistic theory of this type, my best guess is that artificial consciousness will arrive relatively soon, within the next century, and that even farther down the road people will be able to migrate their minds to new hardware much like we now migrate essential data and algorithms from an obsolete computer to an upgraded model. That type of technology will obviously be transformational, though whether good or bad I am not sure. Every aspect of human existence—culture, politics, health, preservation of knowledge and wisdom across periods of time, human dispersion across space, and other environments hostile to biology—will be fundamentally changed by the easy transferability of minds to new hardware. As crazily science fiction as these possibilities sound, I see our technology moving in that direction. My hope is that AST will provide some initial insights into consciousness that are concrete enough, and mechanistic enough, that engineers can build upon it to facilitate the technology.

THE CRUCIAL DIFFERENCE BETWEEN MIND AND LAPTOP

Before explaining the theory, it is useful to specify what phenomenon it purports to tackle. The term consciousness, after all, has many, sometimes conflicting meanings. To help specify the meaning used here, consider the difference between a brain and a modern personal computer. Of course there are many differences, but one seems more consequential than others. The brain has a subjective experience associated with a subset of the information that it processes.

You can connect a computer to a camera and program it to process visual information—color, shape, size, and so on. The human brain does the same, but in addition, we report a subjective experience of those visual properties. This subjective experience is not always present. A great deal of visual information enters the eyes, is processed by the brain and even influences our behavior through priming effects, without ever arriving in awareness. Flash something green in the corner of vision and ask people to name the first color that comes to mind, and they may be more likely to say "green" without even knowing why. But some proportion of the time we also claim, "I have a subjective visual experience. I *see* that thing with my conscious mind. Seeing *feels* like something." The same kind of subjective experience can pertain to other sensory events—a sound, a touch, heat and cold, and so on.

Consider another domain of information: episodic memory. It is a part of our self-identity. It provides a sense of a trajectory through life. But memory itself is not fundamentally mysterious. A computer can store memory, including elaborate information about its past states. Those memories can be retrieved and used to guide output. The crucial, human difference is not that we have memories, or that we can recall them, but that we have a subjective *experience* of memories as we recall them.

Consider one more information-processing event: a decision. Once more, decision-making is not fundamentally mysterious. A computer can make a decision. It can take in information, integrate it, and use it to select one course of action out of many. The human brain also makes decisions. Most of those decisions, possibly tens of thousands a day, occur automatically with no subjective experience, much like in a computer. Yet in some instances, we also report a subjective awareness of making the decision. We sometimes call it intention, choice, or free will. The ability to make a decision, in itself, is not a special human capability. The crucial difference between a personal computer and a human brain lies in the subjective experience that is, sometimes, associated with decision-making—or with memory, sensory processing, or other events in the brain.

This subjective experience is often called consciousness. I admit the term can be misleading. To some people, consciousness refers to a metaphysical soul that floats free of the body after death. To many people it refers to the rich contents swirling within a mind. To some it refers specifically to the part inside you that has free will and chooses one action over another. I mean none of these things. I am referring to the human claim that we have a subjective experience of anything at all. In this account, I will use the terms consciousness, subjective awareness, and subjective experience interchangeably, to refer to this phenomenological property that people claim is associated with some select events and information in the brain.

Like many scientists who study consciousness, I focus on a microcosmic problem: a person looking at a small round spot on a screen (e.g., Webb et al., 2016a). In some circumstances, the person could say, "I have a subjective experience of seeing that spot." In other circumstances, the spot is processed by the visual system, has a measurable impact on the person, and even affects the person's speech and decisions, and yet the person will report, "I didn't consciously see anything." What is the difference between these two circumstances? Why is subjective awareness attached to the visual event in one case and not the other? If we can understand the relevant brain processes for awareness of a spot on a screen, then in principle we can extend the explanation to any information domain. We would understand how people have a subjective experience of vision, touch, sound, the internal richness of memory, mental imagery, decision-making, and self. We would understand the conscious mind.

My point here is that most of what composes the conscious mind is, in principle, not a fundamental mystery. What has resisted explanation thus far is not the content of our experience, but the presence of subjective experience itself. I argue that subjective experience is a confined, relatively easy piece of the neural puzzle to solve.

I also argue that the solution is no mere philosophical flourish. Instead, it is a crucial part of the way the system models and controls itself. It is a key part of the engineering. Without understanding the subjective awareness piece, it may be impossible to build artificial intelligence that has a human-like ability to focus its computational resources and intelligently control that focus. It may also be impossible to build artificial intelligence that can interact with people in a socially competent manner. The study of consciousness is sometimes mistaken as a pursuit of metaphysical mystery, without any practical consequences. The AST does not address a metaphysical mystery. It addresses a concrete piece of the neural puzzle, as pragmatic as the transmission mechanism in a car.

GRASPING AN APPLE WITH THE HAND

The idea of an attention schema was developed in analogy to the body schema. The body schema is an internal model, a rich and integrated set of information that reflects the state of the body, how it moves, and its relationship to the world (Head and Holmes, 1911; Shadmehr and Mussa-Ivaldi, 1994; Graziano et al., 2000; Graziano and Botvinick, 2002; Holmes and Spence, 2004). The body schema not only contributes to the brain's control of the body but also contributes to cognition and verbal behavior. It allows the brain to draw conclusions and make claims about the body. Without a body schema, we would not know that we have a body—except in an intellectual sense, the same way we all know that we have a pancreas. With a body schema, we report having whatever shape or type of body is represented by that body schema. The present section describes the body schema and some of its implications. The following section will draw parallels to an attention schema and our claim to have awareness.

To understand the body schema, consider the body as a robotic device (it could be legitimately called a biological robot) and the brain as the information processor that controls it. Suppose this robot has reached out and grasped an apple. We want to know what information is available to that robot's brain. Three specific types of information are relevant to this discussion: information about the apple, about the robot's own body, and about the physical relationship between the robot and the apple. One of the most important and overlooked aspects of the body schema is that it is not just a representation of the body itself. It contains information about the relationship between the body and the rest of the world.

We will begin with the apple. We ask this biological robot what it is holding, and the robot answers, "An apple." We ask the robot, "Can you describe the apple?" and the robot does so. How does the robot do this? Its brain contains linguistic and cognitive machinery. The cognitive machinery has partial access to the models constructed within its visual system. Its visual system has constructed a rich model of the apple, a set of information about size, color, shape, location, and other attributes, constantly updated as new signals are processed. Due to the presence of this information, and due to the cognitive and linguistic access to the information, the machine is able to respond. It is worth noting that the robot is not actually telling you about the apple. It is telling you about the model of an apple, essentially a simulation, constructed in its visual system. If the internal model contains an error, if it represents the apple as twice too big, for example, the machine will report that incorrect information.

Next, we ask the robot, "What is the state of your body?" Once again, the robot can answer. The reason is that the brain has constructed a body schema—a set of information, constantly updated as new signals are processed, that specifies the size and shape of the limbs and torso and head, how they are hinged, the state they are in at each moment, and what state they are likely to be in over the next few moments. The primary purpose of a body schema is to allow the brain to control movement. A secondary consequence of the body schema is that the robot can explicitly talk about its body. Its cognitive and linguistic processors have some access to the body schema, and therefore the robot can describe its physical self.

Once again, it is worth noting that the robot is not reporting on the actual state of its body, but rather reporting the contents of an internal model. If that internal model is in error, then the robot will provide an incorrect report. If you trick the body schema into representing the arm as more to the left than it actually is, or larger than it actually is, that distorted information will pass through cognition and linguistic processing and enter the verbal report. Even rather extreme illusions of the body schema are easily induced, such as the rubber hand illusion (Botvinick and Cohen, 1998) or the Pinocchio illusion (Lackner, 1988). It is also worth noting that even when the body schema is working correctly, it is always incomplete. It does not contain information about, for example, bone structure, tendon attachments, or the biophysics of muscle contraction. Our biological robot cannot access its body schema and on that basis tell you about the actin and myosin fibers in the muscles. Its body schema contains only the information that the system needs to control the body. The body schema is, in a sense, a cartoon sketch of the body.

Finally, we ask the robot, "What is your physical relationship to the apple?" The robot says, "My arm is outstretched and my hand is grasping the apple." The answer requires integrating two different internal models: the visual system's model of the apple and the body schema. The machine has constructed an amazingly complex, brain-spanning meta-model. Yet in its essence, the behavior remains simple. The machine constructs internal models descriptive of its world. It can report the information content of those internal models because its cognitive and linguistic mechanisms have at least partial access to those internal models. Nothing here is mysterious. Nothing is outside the realm of engineering. I argue that the biological robot, as described thus far, could be copied in artificial form using today's engineering expertise, and it would function in essentially the same way.

I use the term "robot" to communicate a mechanistic perspective, but I intend to describe a human being. We operate in the manner described above. If you hold an apple, the reason why you can say so is that your brain has constructed an internal model of the apple and of your body, integrated those two models to form a larger, overarching description of your physical relationship to that apple, and cognitive and linguistic machinery has access to those internal models. There is something tautological about my central assertion: every claim a person makes, even a simple claim like, "Right now I'm holding an apple," depends on information constructed in the brain. Without the requisite information, the system would be unable to make the claim.

GRASPING AN APPLE WITH THE MIND

Suppose the robot as described above is asked another question. We ask it, "What is the mental relationship between yourself and the apple?" If the robot contains only an internal model of the apple and of a body schema, I argue that it would not be able to answer the new question. It would lack sufficient information. It has sufficient information to answer basic questions about its physical body, about the apple, and about the physical relationship between the two. But a mental relationship? It lacks information on what a mental relationship is. We could ask, "Are you conscious of the apple?" but given the information present, the machine could provide only concrete and literal information such as, "There is an apple." We could press and say, "Yes, but do you have an internal, subjective experience of it?" How could the machine answer? Thus far, we have not given it information to process that question. It would be like asking a digital camera whether it is aware of the picture it just took. The question is meaningless.

Almost all theories of consciousness focus on how a brain might generate a feeling of consciousness. The AST takes a more pragmatic approach, asking how a machine can make the claim that it has a subjective experience. It is a theory about how the brain constructs the requisite information such that the person can make that specific claim. Without the requisite information, the claim cannot be made.

The AST is, in a sense, a proposed extension of the body schema. The proposal is that the brain constructs not only a model of the physical body but also a model of its own internal, information-handling processes. It constructs an "attention schema." That attention schema not only contributes to the control of attention but the information contained within it also has consequences for the kinds of claims that the machine can make about itself.

Attention is a catchall term that arguably adds more confusion than clarity, given its many connotations and meanings. Here, I will mainly avoid the term and use the phrase, "enhanced processing." I will occasionally use the term "attention" when nothing else captures the intended meaning succinctly. The phenomenon I outline below matches at least some uses of the term attention, especially as described by the neuroscientific, "biased competition" theory of attention (Desimone and Duncan, 1995; Beck and Kastner, 2009).

Signals in the brain can be selectively enhanced. For example, consider again the robot from the previous section that encounters an apple. Its visual system constructs a representation of the apple. Under some circumstances, that representation may be suppressed in favor of other representations. Perhaps a sandwich, or another person, or something startling like a bear, wins a competition of visual signals, rises in signal strength, and suppresses the representation of the apple. Under other circumstances, the apple becomes the focus of processing and its representation is enhanced at the expense of other visual representations. This constantly shifting competition among signals can be slanted or biased toward one item or another by a variety of influences, including bottom-up influences (such as a suddenly moving object that causes a surge of signal in the visual system) or top-down influences (such as a cognitive decision to focus one's resources on a specific task). If the apple's representation in the visual system gains in signal strength, winning the competition of the moment, that enhanced processing has a suite of consequences. The apple is processed in greater depth—its nuances and details are more fully processed. It is also more likely to affect other systems throughout the brain, beyond the visual system. The signal is, in effect, broadcasted to other brain areas. It is therefore more likely to affect behavioral decision-making. Whether you reach for the apple or not, bite it, put it away, or decide not to touch it because it looks rotten, the processing of the apple has an impact on behavioral choice. The apple is also more likely to impact memory, allowing it to be recalled later and affect future behavior.

The focusing of resources described here is not limited to a spatial focus. One can focus processing resources on color, on motion, on a particular shape, or on other non-spatial features. It is also not limited to vision. The same type of selective, enhanced processing can be seen in audition, touch, and presumably smell and taste. One can apply the same enhanced processing to movement commands during a difficult movement sequence. It is even possible to selectively enhance entirely internal signals, such as recalled memories, visual imagination, or internal speech. The constantly shifting, enhanced processing of some signals over others, across a vast range of information domains, is one of the most fundamental attributes of the brain.

Now consider again the robot holding an apple. Suppose the machine is focusing its processing resources on the apple. You ask the robot, "What is your mental relationship to the apple?" Can the robot answer this question? Does it have sufficient internal information to report what it is doing computationally? According to AST, the robot can indeed answer the question, and the reason is that it contains an attention schema. The attention schema is a set of information that describes the act of focusing resources on something. The attention schema describes what attention is, what it does, what its most basic stable properties are, what its dynamics and consequences are, and monitors its constantly changing state. Given the information in the attention schema, and given cognitive and linguistic access to at least some of that information, the machine is able to say, "I have a mental grasp of the apple."

Just as the body schema lacks information about mechanistic details such as bone structure and tendon insertion points, so the proposed attention schema lacks detailed information about how signals in the brain are selectively enhanced. The proposed attention schema lacks information about neurons, synapses, electrochemical signals, neural competition, and so on. It has a relatively impoverished description. Suppose you ask the machine, "Tell me more about this mental possession. What physical properties does it have?" The machine is not going to be able to give a scientifically accurate answer. It cannot describe the neuroscience of attention. It replies on the basis of the information available in the attention schema. It says, "My mental possession of that apple, the mental possession in and of itself, has no describable physical properties. It just is. It's a non-physical part of me. My arms and legs are physical parts of me; they have substance. Whatever's inside me that has mental possession of things, that part is non-physical. It's metaphysical. It's my awareness."

It is important to point out what I am not saying. It is easy to imagine building a machine that says, "I am aware of the apple." Just record that message on your phone, then press play, and the machine will utter the phrase. That superficial solution is not what is being described here. What is crucial here is the presence of a rich, descriptive model that is constructed beneath the level of cognition and language, and yet still is accessible to cognition. Because the machine is responding on the basis of an internal model, the response can be flexible, self-consistent, and meaningful. If you ask the machine for more details, it can give a rich description. It might add, "That non-physical, subjective part of me, the real me, is located inside my body. It hovers in my head. It's more or less vivid depending on circumstances. Now that I'm aware of that apple, I *know* about it, what it is and what it's good for. I can choose to react to it. I'll be able to remember it for later. Those are just some of the consequences of awareness. And awareness is not limited to apples. I sometimes experience other things as well. Right now I'm aware of you, sometimes I experience a flood of recalled memories, or mental imagery that I invent fancifully, and sometimes I have the subjective experience of making a decision. There's a commonality across all those circumstances—I have a subjective, mental possession of things inside me and around me." In this description, the machine is coming close to the literal truth. It is giving a fairly close, if high-level and detail-poor, description of how it focuses its processing resources on one or another item. Its description veers from literal reality only as it muddles the more mechanistic details and ultimately claims to have a spooky, physically incoherent consciousness. Consciousness is, in a sense, a cartoon sketch of attention.

Suppose you ask the machine, "But aren't you making all those claims simply because that's the information contained in your internal models? Aren't you just a computing machine?"

The machine accesses its internal models and finds nothing to match your suggestion. Its internal models do not announce to cognition, "By the way, this is information contained in an internal model, and the information might not be literally accurate." On the basis of the limited information available, the machine says, "What information? What internal models? This has nothing to do with computation. No, I am simply subjectively aware of the apple." The machine is captive to its own information. It knows only what it knows.

Colleagues have often asked me: granted that the brain probably does construct something like an attention schema, how does that internal model explain how we have subjective experience? Why does it *feel* like anything at all to process information? The answer is that the theory emphatically does not explain how we have a subjective experience. It explains how a machine *claims* to have a subjective experience, and how it is that the machine cannot tell the difference.

The AST has some similarities to the illusionist approach to consciousness (e.g., Dennett, 1991; Norretranders, 1999; Frankish, 2016). In that view, subjective experience is not truly present; instead, the brain is an entirely mechanistic processor of information that has an illusion of possessing consciousness. Exactly how the illusion occurs differs somewhat between accounts. Clearly, the illusionist approach has a philosophical similarity to the AST. However, I remain uncomfortable with calling consciousness an illusion. In AST, the brain does not experience an illusion. It does not subjectively experience anything. Instead, the machine has wrong, or simplified information that tells it that it is having an experience. In my view, calling consciousness an illusion is trying too hard to employ an everyday, intuitive concept that is not truly applicable.

Another similar approach to consciousness might be called the "naïve theory" perspective (e.g., Gazzaniga, 1970; Nisbett and Wilson, 1977; Dennett, 1991). In that view, the brain processes information about its world but does not possess any subjective experience. We claim that we do because, at a cognitive level, we have learned a naïve theory. It is essentially a ghost story, a socially learned narrative that we use to explain ourselves, a social epiphenomenon with debatable utility. With different upbringing, we would not claim to have any conscious experience. Again, there is some philosophical similarity between this view and AST. Indeed, the two are very close. However, in AST, the naïve construct of consciousness is not learned. It is not at a higher cognitive level. It is wired into the system at a deep level and constructed automatically, like the body schema. It is inborn. As discussed below, it is probably present in a range of species. Moreover, it is not a social epiphenomenon; instead, it serves a specific set of important cognitive functions. The brain constructs internal models because of the specific usefulness of modeling and monitoring items in the real world, and the usefulness of the attention schema is the crux of the theory, as discussed in the following sections.

The AST also has strong similarities to approaches in machine consciousness (e.g., Chella et al., 2008) in which a system can contain representations of the self, the environment, and higher order, recursive representations of how the self relates to the environment. This general concept resonates closely with the concepts of the AST. The AST is a theory of how the human brain models its own human-like attention systems and thus makes the claim that it has a subjective experiential component. Artificial systems that have different internal architecture, perhaps different processes akin to but not identical to human attention, might require different self-representations. A machine of that nature would not necessarily lay claim to consciousness in the sense that we humans intuitively understand it. Drawing on its own internal quirky representations, it would describe itself in ways specific to it. Of course, we might expect the contents of that machine's mind to differ from a human's mind. But, the point I am trying to make here is that the very construct of consciousness, of subjective experience itself, whether the machine even has that construct and what the details of it may be, will depend on the precise nature of the machine's internal models.

THE ADAPTIVE VALUE OF AN ATTENTION SCHEMA: CONTROL OF ATTENTION

The sections above discussed the consequences of cognitive and verbal access to internal models. For example, the body schema allows you to close your eyes and still know about and talk about the configuration of your body. The primary function of the body schema, however, is probably less for cognitive access and more for the control of movement. One of the fundamental principles in control engineering is that a good controller contains a model of the item being controlled (Conant and Ashby, 1970; Francis and Wonham, 1976; Camacho and Bordons Alba, 2004; Haith and Krakauer, 2013). A robot arm, the airflow throughout a building, a self-driving car, each system benefits from an appropriate internal model. The model partly monitors the state of the item to be controlled and also partly predicts states into the near future. The body schema contains layers of information about the body, about its stable properties such as its shape and hinged structure and about more dynamic properties such as forces and velocities (Head and Holmes, 1911; Shadmehr and Mussa-Ivaldi, 1994; Shadmehr and Moussavi, 2000; Graziano and Botvinick, 2002; Holmes and Spence, 2004; Hwang and Shadmehr, 2005). This information is used during the control of movement for obstacle avoidance, for on-line error correction, and for longer term adaptation. If movements are systematically wrong or distorted, the internal model can be adapted to correct the errors.

We hypothesized that the same advantages accrue from having an attention schema. The ability to focus processing resources strategically on one or another signal requires control. That control should benefit from an attention schema—a coherent set of information that represents basic stable properties of attention, reflects ongoing changes in the state of attention, makes predictions about where attention can be usefully directed, and anticipates consequences of attention. The best way to test this hypothesis would be to isolate cases where awareness fails—cases where the brain is processing information but people report being unaware of it. In those cases, by hypothesis, the attention schema has failed. While the system may still be capable of directing attention, focusing resources on the signal in question, the control of attention should suffer in characteristic ways—much like the control of the arm might become more wobbly, less able to error-correct, and less adaptable over repeated trials, if the arm's internal model is compromised.

Several experimental results on attention and awareness have been interpreted as consistent with this prediction (McCormick, 1997; Tsushima et al., 2006; Lin and Murray, 2015; Webb and Graziano, 2015; Webb et al., 2016a), though more experiments are needed. Thus far, the relevant experiments have focused on visual attention and visual awareness. When people are unaware of a visual stimulus, they can still sometimes focus processing resources on it. They can direct attention to it (McCormick, 1997; Lamme, 2003; Woodman and Luck, 2003; Ansorge and Heumann, 2006; Tsushima et al., 2006; Kentridge et al., 2008; Hsieh et al., 2011; Norman et al., 2013). However, in that case, visual attention suffers deficits in control. It behaves less stably over time and shows evidence of being less able to error-correct and less able to adapt to perturbations (McCormick, 1997; Lin and Murray, 2015; Webb and Graziano, 2015; Webb et al., 2016a). The evidence suggests that awareness is necessary for the good control of attention.

One group of researchers has presented a computational model of attention with and without an internal model and found that at least this simplified, artificial attention is better controlled with the internal model (van den Boogaard et al., 2017).

In our hypothesis, the attention schema first evolved as a crucial part of the control system for attention. The possible co-evolution of attention and awareness has been discussed before (Graziano, 2010, 2013, 2014; Haladjian and Montemayor, 2015; Graziano and Webb, 2016). Since the basic vertebrate brain mechanisms for controlling attention emerged more than half a billion years ago, we speculate that the origin of awareness, at least in preliminary form, may be equally ancient. Awareness, in this view, is not simply a philosophical flourish. It is a part of the engineering. Just as one cannot understand how the brain controls the body without understanding that the brain constructs a body schema, so one cannot understand how the brain intelligently deploys its limited processing resources without understanding that it constructs an attention schema. That an attention schema causes us humans to lay claim to a metaphysical soul is a quirky side effect.

THE ADAPTIVE VALUE OF AN ATTENTION SCHEMA: SOCIAL COGNITION

One of the most devastating impairments to awareness in the clinical literature is hemispatial neglect. Damage to one side of the brain, typically the right temporoparietal junction (TPJ), causes a loss of awareness of everything to the opposite side of space (Vallar and Perani, 1986; Corbetta, 2014). Yet, information from the neglected side is still processed to some degree (Marshall and Halligan, 1988), and the visual system is still active to the highest levels of processing (Rees et al., 2000; Vuilleumier et al., 2002). Neglect appears to be caused by the disruption of brain networks involved in attention and awareness that pass through the TPJ (Corbetta, 2014; Igelström and Graziano, 2017).

The TPJ, however, has also been implicated in social cognition. When people attribute mind states to each other, such as beliefs or emotions, brain-wide networks are recruited that also pass through the TPJ (Saxe and Wexler, 2005; Kelly et al., 2014; Igelström et al., 2016). A complicated literature suggests that, although there is some separation of function among subregions of the TPJ, considerable overlap of function is also present (Mitchell, 2008; Scholz et al., 2009; Igelström et al., 2016; Igelström and Graziano, 2017). The adjacency and possible overlap of social cognition functions with awareness and attention functions has caused some controversy.

We suggested that the functional overlap within the TPJ may have a deeper significance (Graziano and Kastner, 2011; Graziano, 2013). In our proposal, one of the primary uses for the construct of awareness is for social cognition. We attribute to other people an awareness of the objects and events around them. When we do so, we are in effect constructing a simplified model of other people's state of attention. Arguably, all of social cognition depends on attributing awareness to other people. Does Frank intend to walk toward you, or sit in that chair, or eat that sandwich? Only if he is aware of you, the chair, or the sandwich. Is he angry that someone made a rude gesture at him? Only if he is aware of the gesture. Whether reconstructing someone else's beliefs, intentions, emotions, or any other mental state, we depend first on attributions of awareness.

In our hypothesis, the TPJ is a central node in a brain-wide network that helps to compute an attention schema. That attention schema is our construct of awareness, and that construct can be applied to oneself or to others. Much like the color-processing networks in the visual system can assigned colors to surfaces, so the social cognition network can assign the construct of awareness to agents, including oneself. Experimental evidence from brain imaging studies suggests that the TPJ does play a role in attributing visual awareness to others, and that some of the same subregions of the TPJ are involved in constructing one's own visual awareness (Kelly et al., 2014; Igelström et al., 2016; Webb et al., 2016b). We suggest that the TPJ is a site where the ability to perceive consciousness in others grew out of our ability to be conscious ourselves. However, the TPJ remains an extremely complex area of the cortex that is still poorly understood. Far more work will be needed to specify its range of functions and how they are distributed anatomically.

Given the goal of this article, introducing AST to those who may be interested in engineering it, the specific networks in the brain are not of great importance. Whether the computations are performed by this or that part of the brain are irrelevant. What is important is the overlap in function between modeling oneself and modeling others. A mechanism that can compute an internal model of attention, an attention schema, may be important not just for controlling one's own attention, but also for monitoring the attentional states of others. The social use of an attention schema may be especially developed in humans. We attribute awareness to each other, to pets, to inanimate objects, and to the spaces around us. Arguably, the entire spirit world, from deities down to minor ghosts, owes itself to our social neural machinery building the construct of awareness and attributing it promiscuously to ourselves and everything else around us. To build machines with similar social ability, the ability to attribute consciousness to itself and to others, such that the machine can understand what it means for another agent to be conscious, may require something like an attention schema.

WHY BUILD ARTIFICIAL CONSCIOUSNESS?

If AST is correct, then consciousness is buildable with current technology. In this respect, the theory differs from other major theories of consciousness that provide much less clear direction for how to build consciousness.

For example, the global workspace theory posits that the brain-wide boosting and broadcasting of a signal, such as a visual signal, causes that signal to enter consciousness (Baars, 1988; Dehaene, 2014). In effect, the global workspace theory is the same as the AST, if you took away the attention schema part, and had only the attention part—the ability of the brain to selectively enhance signals such that they have a global impact on many brain systems. While in my view the theory is likely to be correct as far as it goes, it is incomplete. It does not explain why the globally broadcasted information would be associated with the property of subjective experience. Building a machine that has signals boosted in that manner, to a strength sufficient to globally effect other systems in the machine, is easily done and arguably has already been done. But it is not a good prescription for building consciousness. There is no reason to suppose that a machine of that sort would sit up and say, "Wow, I have an internal experience of these things." It brings us no closer to the behavior that humans exhibit, namely, claiming to have subjective awareness.

The integrated information theory (Tononi, 2008) suffers a similar problem. In that theory, consciousness is the result of highly integrated information in the brain. A mathematical formula can tell you how much integrated information, and thus how much consciousness, is present in any specific device. To many scientists, including myself, this theory is non-explanatory and ultimately unfalsifiable. It is somewhat like the science fiction trope: if you build a computer big and complex enough, integrating enough information together, it will somehow become conscious. To be fair to the theory, in my view, there is likely to be at least some type of relationship between consciousness and highly integrated information. Even in AST, the proposed attention schema is a bundle of information that is integrated with other schemas and models around the brain. But as a prescription for building consciousness, the integrated information theory by itself has been disappointing, since even very complex technology that contains a lot of integrated information has not announced its consciousness yet.

The AST instead presents an extremely simple conceptual foundation. The machine claims to be conscious of items and events, because it constructs information that describes that condition of consciousness. Without the internal information indicating that it contains consciousness, it would not be able to make the claim. The reason why it constructs that quirky internal information is because it is a useful, if not literally accurate, model of the machine's ability for deep, focused processing. The AST therefore points a practical way toward building a machine that makes the same claims of consciousness that people do.

I recognize that AST is not yet specific enough to hand a blueprint to an engineer. Yet, it lays a conceptual foundation for building consciousness. Because it is a theory in which a machine constructs a specific set of information and uses it in a specific way, it is buildable. Given current technology, an enterprising set of AI researchers should be able to build a machine that contains a fairly rich model of what consciousness is and that can attribute the property of consciousness to itself and to the people it interacts with. It should be possible to build a machine that believes it is conscious and claims it is conscious and acts like it is conscious and that talks about its consciousness in the same ways that the human machine does.

Why try to build artificial consciousness? One could build it for entertainment value. It would be monumentally cool. But I also see two practical reasons. The first may be of technical interest to specialists, whereas the second is of fundamental importance to all of us.

First, evolution has given us effective brains, and copying the biological solution might make for capable artificial intelligence. Suppose that the theory is correct, and consciousness depends on an attention schema. With an attention schema acting as an internal control model, the brain is better able to control and deploy its limited processing resources. Perhaps giving machines a human-like focus of attention, and an attention schema, will be helpful. Artificial systems might thereby become better able to control their own limited processing resources. Admittedly, I do not know if this engineering trick borrowed from the brain will be of use to artificial intelligence. Computer systems can process more information, more quickly, than biological systems, and can be organized in fundamentally different ways. It is not clear whether human-like attention, or human-like control of attention, would necessarily benefit artificial systems. The idea would be worth pursuing, but better engineering solutions might be discovered along the way.

To me the most compelling reason to pursue artificial consciousness is that, if the theory is correct, then consciousness is the foundation of social intelligence. An agent cannot be socially competent unless it has a fairly rich internal model of what consciousness is and can attribute consciousness to itself and to other people. If we want to build machines that are skilled at interacting with people, we will need to build in consciousness in the same sense that people attribute consciousness to themselves and see consciousness in others. It is the root of empathy. Without that capacity, our computers are sociopaths. A similar point has been made by others, including the point that social capability is urgently needed in artificial intelligence (e.g., Sullins, 2016), and that self-models are a crucial part of human social competence (e.g., Hood, 2012).

While human sociopaths are evidently conscious—they can attribute that property to themselves—they are impaired at attributing it to others. They may know intellectually that other people contain minds, but they appear to lack a fundamental, automatic perception of the consciousness of others. Other people are mechanical objects to them. Half of the functional range of the attention schema is impaired. We cannot build machines that treat people with humanistic care, if they do not have that crucial social capability to attribute consciousness to others. Machine consciousness is a necessary step for our future. For those who fear that AI is potentially dangerous and may harm humanity, I would say that the danger is infinitely greater with sociopathic computers and it is of the utmost priority to give them consciousness—both the ability to attribute it to themselves and to others. I urge anyone with the technical expertise, who is reading this article, to think about how to tackle the problem.

AUTHOR CONTRIBUTIONS

MG is responsible for all aspects of this article.

FUNDING

This study was funded by the Princeton Neuroscience Institute Innovation Fund.

REFERENCES

Ansorge, U., and Heumann, M. (2006). Shifts of visuospatial attention to invisible (metacontrast-masked) singletons: clues from reaction times and event-related potentials. *Adv. Cogn. Psychol.* 2, 61–76. doi:10.2478/v10053-008-0045-9

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. *Vis. Res.* 49, 1154–1165. doi:10.1016/j.visres.2008.07.012

Botvinick, M., and Cohen, J. D. (1998). Rubber hand 'feels' what eye sees. *Nature* 391, 756. doi:10.1038/35784

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

Chella, A., Frixione, M., and Gaglio, S. (2008). A cognitive architecture for robot self-consciousness. *Artif. Intell. Med.* 44, 147–154. doi:10.1016/j.artmed.2008.07.003

Conant, R. C., and Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *Int. J. Syst. Sci.* 1, 89–97. doi:10.1080/00207727008920220

Corbetta, M. (2014). Hemispatial neglect: clinic, pathogenesis, and treatment. *Semin. Neurol.* 34, 514–523. doi:10.1055/s-0034-1396005

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

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

Desimone, R., and Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annu. Rev. Neurosci.* 18, 193–222. doi:10.1146/annurev.ne.18.030195.001205

Francis, B. A., and Wonham, W. M. (1976). The internal model principle of control theory. *Automatica* 12, 457–465. doi:10.1016/0005-1098(76)90006-6

Frankish, K. (2016). Illusionism as a theory of consciousness. *J. Conscious. Stud.* 23, 11–39.

Gazzaniga, M. S. (1970). *The Bisected Brain*. New York: Appleton Century Crofts.

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. *J. Cogn. Neurosci.* 26, 1300–1304. doi:10.1162/jocn_a_00623

Graziano, M. S. A., and Botvinick, M. M. (2002). "How the brain represents the body: insights from neurophysiology and psychology," in *Common Mechanisms in Perception and Action: Attention and Performance XIX*, eds W. Prinz and B. Hommel (Oxford: Oxford University Press), 136–157.

Graziano, M. S. A., Cooke, D. F., and Taylor, C. S. R. (2000). Coding the location of the arm by sight. *Science* 290, 1782–1786. doi:10.1126/science.290.5497.1782

Graziano, M. S. A., and Kastner, S. (2011). Human consciousness and its relationship to social neuroscience: a novel hypothesis. *Cogn. Neurosci.* 2, 98–113. doi:10.1080/17588928.2011.565121

Graziano, M. S. A., and Webb, T. W. (2016). "From sponge to human: the evolution of consciousness," in *Evolution of Nervous Systems*, 2nd Edn, Vol. 3, ed. J. Kaas (Oxford: Elsevier), 547–554.

Haith, A. M., and Krakauer, J. W. (2013). "Model-based and model-free mechanisms of human motor learning," in *Progress in Motor Control: Advances in Experimental Medicine and Biology*, Vol. 782, eds M. Richardson, M. Riley, and K. Shockley (New York: Springer), 1–21.

Haladjian, H. H., and Montemayor, C. (2015). On the evolution of conscious attention. *Psychon. Bull. Rev.* 22, 595–613. doi:10.3758/s13423-014-0718-y

Head, H., and Holmes, G. (1911). Sensory disturbances from cerebral lesions. *Brain* 34, 102–254. doi:10.1093/brain/34.2-3.102

Holmes, N., and Spence, C. (2004). The body schema and the multisensory representation(s) of personal space. *Cogn. Process.* 5, 94–105. doi:10.1007/s10339-004-0013-3

Hood, B. (2012). *The Self Illusion: How the Social Brain Creates Identity*. New York: Oxford University Press.

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. *Psychol. Sci.* 22, 1220–1226. doi:10.1177/0956797611419302

Hwang, E. J., and Shadmehr, R. (2005). Internal models of limb dynamics and the encoding of limb state. *J. Neural Eng.* 2, S266–S278. doi:10.1088/1741-2560/2/3/S09

Igelström, K. M., and Graziano, M. S. A. (2017). The inferior parietal lobule and temporoparietal junction: a network perspective. *Neuropsychologia*. 105, 70–83. doi:10.1016/j.neuropsychologia.2017.01.001

Igelström, K., Webb, T. W., and Graziano, M. S. A. (2016). Functional connectivity between the temporoparietal cortex and cerebellum in autism spectrum disorder. *Cereb. Cortex* 27, 2617–2627. doi:10.1093/cercor/bhw079

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. *Proc. Natl. Acad. Sci. U.S.A.* 111, 5012–5017. doi:10.1073/pnas.1401201111

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. doi:10.1016/j.neuropsychologia.2007.11.036

Lackner, J. R. (1988). Some proprioceptive influences on the perceptual representation of body shape and orientation. *Brain* 111, 281–297. doi:10.1093/brain/111.2.281

Lamme, V. A. (2003). Why visual attention and awareness are different. *Trends Cogn. Sci.* 7, 12–18. doi:10.1016/S1364-6613(02)00013-X

Lin, Z., and Murray, S. O. (2015). More power to the unconscious: conscious, but not unconscious, exogenous attention requires location variation. *Psychol. Sci.* 26, 221–230. doi:10.1177/0956797614560770

Marshall, J. C., and Halligan, P. W. (1988). Blindsight and insight in visuo-spatial neglect. *Nature* 336, 766–767. doi:10.1038/336766a0

McCormick, P. A. (1997). Orienting attention without awareness. *J. Exp. Psychol. Hum. Percept. Perform.* 23, 168–180. doi:10.1037/0096-1523.23.1.168

Mitchell, L. P. (2008). Activity in the right temporo-parietal junction is not selective for theory-of-mind. *Cereb. Cortex* 18, 262–271. doi:10.1093/cercor/bhm051

Nisbett, R. E., and Wilson, T. D. (1977). Telling more than we can know – verbal reports on mental processes. *Psychol. Rev.* 84, 231–259. doi:10.1037/0033-295X.84.3.231

Norman, L. J., Heywood, C. A., and Kentridge, R. W. (2013). Object-based attention without awareness. *Psychol. Sci.* 24, 836–843. doi:10.1177/0956797612461449

Norretranders, T. (1999). *The User Illusion: Cutting Consciousness Down to Size*. New York: Penguin.

Rees, G., Wojciulik, E., Clarke, K., Husain, M., Frith, C., and Driver, J. (2000). Unconscious activation of visual cortex in the damaged right hemisphere of a parietal patient with extinction. *Brain* 123, 1624–1633. doi:10.1093/brain/123.8.1624

Saxe, R., and Wexler, A. (2005). Making sense of another mind: the role of the right temporo-parietal junction. *Neuropsychologia* 43, 1391–1399. doi:10.1016/j.neuropsychologia.2005.02.013

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., and Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLoS ONE* 4:e4869. doi:10.1371/journal.pone.0004869

Shadmehr, R., and Moussavi, Z. M. (2000). Spatial generalization from learning dynamics of reaching movements. *J. Neurosci.* 20, 7807–7815.

Shadmehr, R., and Mussa-Ivaldi, F. A. (1994). Adaptive representation of dynamics during learning of a motor task. *J. Neurosci.* 14, 3208–3224.

Sullins, J. (2016). Artificial phronesis and the social robot. *Front. Artif. Intell. Appl.* 290:37–39. doi:10.3233/978-1-61499-708-5-37

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull.* 215, 216–242. doi:10.2307/25470707

Tsushima, Y., Sasaki, Y., and Watanabe, T. (2006). Greater disruption due to failure of inhibitory control on an ambiguous distractor. *Science* 314, 1786–1788. doi:10.1126/science.1133197

Vallar, G., and Perani, D. (1986). The anatomy of unilateral neglect after right-hemisphere stroke lesions. A clinical/CT-scan correlation study in man. *Neuropsychologia* 24, 609–622. doi:10.1016/0028-3932(86)90001-1

van den Boogaard, E., Treur, J., and Turpijn, M. (2017). "A neurologically inspired neural network model for Graziano's attention schema theory for consciousness," in *International Work Conference on the Interplay between Natural and Artificial Computation: Natural and Artificial Computation for Biomedicine and Neuroscience, Part 1*, 10–21. doi:10.1007/978-3-319-59740-9_2

Vuilleumier, P., Armony, J. L., Clarke, K., Husain, M., Driver, J., and Dolan, R. J. (2002). Neural response to emotional faces with and without awareness: event-related fMRI in a parietal patient with visual extinction and spatial neglect. *Neuropsychologia* 40, 2156–2166. doi:10.1016/S0028-3932(02)00045-3

Webb, T. W., and Graziano, M. S. A. (2015). The attention schema theory: a mechanistic account of subjective awareness. *Front. Psychol.* 6:500. doi:10.3389/fpsyg.2015.00500

Webb, T. W., Kean, H. H., and Graziano, M. S. A. (2016a). Effects of awareness on the control of attention. *J. Cogn. Neurosci.* 28, 842–851. doi:10.1162/jocn_a_00931

Webb, T. W., Igelström, K., Schurger, A., and Graziano, M. S. A. (2016b). Cortical networks involved in visual awareness independently of visual attention. *Proc. Natl. Acad. Sci. U.S.A.* 113, 13923–13928. doi:10.1073/pnas.1611505113

Woodman, G. F., and Luck, S. J. (2003). Dissociations among attention, perception, and awareness during object-substitution masking. *Psychol. Sci.* 14, 605–611. doi:10.1046/j.0956-7976.2003.psci_1472.x

Conflict of Interest Statement: The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.