Routledge
Taylor & Francis Group

Cognitive Neuropsychology

M. S. A. Graziano, A. Guterstam, B. J. Bio, A. I. Wilterson, Toward a standard model of consciousness: reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. Cognitive Neuropsychology 37, 155-172 (2020).

ISSN: 0264-3294 (Print) 1464-0627 (Online) Journal homepage: https://www.tandfonline.com/loi/pcgn20

Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories

Michael S. A. Graziano, Arvid Guterstam, Branden J. Bio & Andrew I. Wilterson

To cite this article: Michael S. A. Graziano, Arvid Guterstam, Branden J. Bio & Andrew I. Wilterson (2019): Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories, Cognitive Neuropsychology, DOI: 10.1080/02643294.2019.1670630

To link to this article: https://doi.org/10.1080/02643294.2019.1670630

COGNITIVE NEUROPSYCHOLOGY
https://doi.org/10.1080/02643294.2019.1670630

Routledge
Taylor & Francis Group

REVIEW

Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories

Michael S. A. Graziano, Arvid Guterstam, Branden J. Bio and Andrew I. Wilterson

Department of Psychology and Princeton Neuroscience Institute, Princeton University, Princeton, NJ, USA

ABSTRACT

Here we examine how people's understanding of consciousness may have been shaped by an implicit theory of mind. This social cognition approach may help to make sense of an apparent divide between the physically incoherent consciousness we think we have and the complex, rich, but mechanistic consciousness we may actually have. We suggest this approach helps reconcile some of the current cognitive neuroscience theories of consciousness. We argue that a single, coherent explanation of consciousness is available and has been for some time, encompassing the views of many researchers, but is not yet recognized. It is obscured partly by terminological differences, and partly because researchers view isolated pieces of it as rival theories. It may be time to recognize that a deeper, coherent pool of ideas, a kind of standard model, is available to explain multiple layers of consciousness and how they relate to specific networks within the brain.

ARTICLE HISTORY
Received 12 March 2019
Revised 12 September 2019
Accepted 16 September 2019

KEYWORDS
Consciousness; awareness; attention; theory of mind; social cognition

Introduction

The ability of people to intuit what might be going on in someone else's head is called theory of mind (Baron-Cohen, 1997; Wellman, 2018; Wimmer & Perner, 1983). In the present article, we discuss how this natural theory of mind may have influenced the study of consciousness—an endeavour which, after all, is a scientific attempt to build a theory of the mind. Some of the most recent data from our lab point toward very strange, irrational, but automatic models of mind, especially of the process of attention, that we all construct at an implicit level (Guterstam, Kean, Webb, Kean, & Graziano, 2018). The inherent inaccuracies and simplifications in this type of self-model may have shaped intuitive convictions, folk beliefs, and even scientific hypotheses, about mind and thinking. At least some of the more mystical, common notions about consciousness may have derived from our imperfect models of mind. This possibility—that people attribute a mysterious consciousness to themselves and to others because of an inherently inaccurate model of mind, and especially a model of attention—was proposed in some detail previously and called the attention schema theory or AST (Graziano, 2013, 2019a, 2019b; Graziano & Kastner, 2011; Webb & Graziano, 2015).

AST is closely related to several other cognitive neuroscience theories of consciousness. In the final section of this article, we argue that these theories should not be viewed as rivals, but as partial perspectives on a deeper mechanism. In particular, the present social cognitive approach may be able to bridge between AST, the global workspace theory (GW), the higher-order thought theory (HOT), and the illusionist perspective. Although many scholars may argue that consciousness remains a mystery and that a single explanation has not yet emerged out of the field of rival explanations, here we argue that neuroscience and psychology have already produced a working theory of consciousness, a kind of standard model that covers the basics if not the details. What stands in the way of a broader recognition of that theory is the fact that different contributions to it have been formulated in different ways, at different times, using different vocabulary, obscuring the deeper underlying convergence of ideas.

i-consciousness and m-consciousness

Studying consciousness scientifically is difficult when the word has multiple meanings. For clarity, we begin this account by labelling two common

CONTACT Michael S. A. Graziano graziano@princeton.edu

categories of meaning for consciousness. The first focuses on information in the brain—how it is selected, enhanced, and processed. The second is a more mysterious, extra, experiential essence that people claim accompanies the informational content. In this account, we will refer to the two as i-consciousness (i for information) and m-consciousness (m for mysterious), although we acknowledge that other researchers may use different terminology. The primary reason for using this two-part terminology is to make it as clear as possible that, in our perspective, at least some form of consciousness exists. We do not argue that consciousness is entirely an illusion or a mistaken construct. Rather, i-consciousness literally and mechanistically exists within us. Somebody is home. We can then debate whether m-consciousness, the more ethereal notion of consciousness that people intuitively believe they have, is accurate or instead is an imperfect model of i-consciousness.

Presently, the most generally accepted theory of i-consciousness is probably the global workspace theory (GW) (e.g., Baars, 1988; Dehaene, 2014; Dehaene & Changeux, 2011; van Vugt et al., 2018). In it, if you look at an object such as an apple, as the visual information is processed in a complex set of brain areas, the signals related to the apple may win an attentional competition, growing in strength and consistency. With sufficient attentional enhancement these signals may reach a threshold where they achieve "ignition", which means that they dominate the larger, brain-spanning networks, especially networks across the parietal and prefrontal cortex. The visual information about the apple becomes available for systems around the brain, such as speech systems that allow you to talk about the apple, motor systems that allow you to reach for it, cognitive systems that allow you to make high-level decisions about it, and memory systems that allow you to store that particular moment for possible later use. In that circumstance, the visual information about the apple has entered the global workspace and thus entered consciousness, whereas the vast majority of other information in the brain has failed to reach the global workspace and thus remained outside of consciousness. In our perspective, though others may disagree, GW is an account of i-consciousness. It is about how select information reaches a state in which it is bundled centrally and can impact output systems. A global workspace is computationally buildable (Baars & Franklin, 2007) and can be studied objectively even in non-human animals performing detection tasks (e.g., van Vugt et al., 2018). One way to summarize the theory is that i-consciousness is associated with the highest levels of attentional enhancement in the cerebral cortex.

Explaining i-consciousness, however, is only part of the challenge. In a traditional perspective, in addition to the content of consciousness, we have something else, something that accompanies the information, or imbues it, or in some manner is the essence of experience. The challenge of explaining consciousness, in this traditional perspective, lies in explaining the extra essence—subjective awareness, or m-consciousness as we label it here. The idea of a distinction between the information contained in consciousness, which can be understood through materialist theories such as GW, and the extra, non-materialist property of subjective experience, emerged mainly during the twentieth century, possibly as a result of the rise of information technology. Without a well-developed concept of an information-processing machine, it is difficult to realize that the information in the mind might be different from the experiential essence of the mind. For example, William James (1890) essentially conflated the two, coining the term "stream of consciousness" to describe the ever-changing mental content. Even at the start of the computer revolution, when Turing (1950) wrote about whether a machine can think, he emphasized the processing of information and not whether the machine can have a subjective experience of that information. But within a few decades, Nagel (1974) argued that it is not enough to process information. There is a non-materialistic component, a "what it is like" to experience something. It is non-physical in the sense that one cannot touch it, weigh it, or snap it in half and measure its tensile strength, but Nagel argued it exists nonetheless. Chalmers (1995) refined the idea, referring to the easy problem of scientifically figuring out how the brain constructs the content of consciousness and the hard problem of figuring out the nature of subjective experience itself. As a result of these ideas, for the past twenty-five years, consciousness has been widely viewed as containing two parts. Whereas i-consciousness can be understood mechanistically and can probably be replicated by machines, m-consciousness is considered difficult or impossible to explain.

Here, we argue that the belief in m-consciousness —the belief in a non-materialistic component to the mind—is a lingering fragment of a larger cluster of physically incorrect beliefs. These beliefs about mind originated deep in the past, may have an evolutionary origin, and may still be present in us at an implicit level, even though modern science has rejected some of them. This cluster of ideas includes the culturally widespread, folk-psychological theory that mind is an energy-like substance inhabiting the body. To be clear, we do not mean to suggest that ghosts exist. Rather, we argue that culturally common beliefs, that derive from implicit, social-cognitive models, have infiltrated the science of consciousness and have led to some mistaken assumptions.

The belief in a spirit world may be as old as the human species. Intentional burial of the dead with grave goods dates back at least 100 thousand years among *Homo sapiens*, and may indicate a belief in a mental, spiritual essence that lives past death (Pettitt, 2010). The monumental stone structures on Gobekli Tepe in modern-day Turkey, built as much as 12 thousand years ago, have been interpreted as sites for communing with a spirit world (Schmidt, 2011). One of the most vivid ancient descriptions of the mind as a gauzy essence separable from the physical body is found in Homer's Odyssey from nearly 3,000 years ago, in the passages in which Odysseus visits the spirits of the underworld. A significant proportion of people still believe in a mental force or energy that is generated inside of a person and flows out of the eyes, touching objects (Gross, 1999; Guterstam et al., 2018; Winer, Cottrell, Gregg, Fournier, & Bica, 2002). Even when people insist they do not believe in such things, their behaviour is measurably affected by what appears to be an implicit belief in eye-beams (Guterstam et al., 2018). Many people still believe in the evil eye, essentially an emotional ill will that can emerge from one person, travel invisibly to another, and negatively impact the recipient (Dundes, 1981). The idea of telekinesis, the supposed ability of a sufficiently focused person to project an invisible force-carrying mental essence and cause objects to move, is also psychologically compelling and culturally widespread (Benassi, Sweeney, & Drevno, 1979). Mesmer's theory of animal magnetism, a special living force field akin to electromagnetism by which we can directly influence each other, had such cultural resonance in the eighteenth century that its echo is still present in our language (Pattie, 1994). The idea of an energy-like spirit resonates so well with general audiences that it is even widespread in fiction, such as the Force in Star Wars, the glowing souls sucked out of people by dementors in the Harry Potter stories (Rowling, 1999), the aura shining from righteous elves in Tolkien's Middle Earth (Tolkien, 1955), or the soul dust in the His Dark Materials series (Pullman, 2000). Given all of these beliefs and tropes across many cultures, it appears that historically, people almost uniformly believed a folk psychological theory in which the mind is an energy-like essence. That essence has at least some physical properties, but lacks others. It typically has no definite weight, size, reflectance, hardness, texture, or many other objectively measurable characteristics. Yet it has a general location inside of a person, it can flow through space especially out of the eyes, it can be directed by effort, it can sometimes have a physical effect on objects external to the agent, and perhaps because of its ability to flow out of the body, it can often survive the death of the body. In that folk perspective, the energy-like mind stuff is the thing with which we experience, think, understand, and actuate our bodies. It is the essence of us. It is our consciousness.

Most scientists and philosophers understandably avoid this "ghost" approach, but its cultural ubiquity may say something important about how people process the world. We suggest that this type of folk psychological belief may derive from deep, automatic models that all people construct as a part of our social cognition. Many people in the modern world believe that they are rid of these ghost beliefs, but we suggest that at an implicit level we all still build ghost-like models of other people's minds (Guterstam et al., 2018). We suggest that people automatically construct a simplified, cartoon version of the social world in which agents—whether others or oneself—possess an invisible, energy-like or plasma-like mental essence. The model contains at least the following components, and probably many more:

(1) A mind is something that can actively hold information, and does so by having a subjective experience.
(2) A mind is a fluidic substance. It has a general location usually inside of an agent, and can

move through space and time, sometimes flowing outside the agent toward objects of attention.

(3) A mind has an energy-like property—it has the ability to physically affect the world. It empowers the agent to act, and also sometimes directly acts on objects surrounding the agent.

In our proposal, this model is, of course, not accurate. Much like the visual system models white light as brightness without any contaminating colors—a simplification of a more complex physical reality—so our intrinsic model of a mind is a cartoonish version of the functioning of an active, attentive brain. And yet, despite the inaccuracy of that model, through most of human history, most people assumed that these spirit-like properties of mind were literally correct (Sidky, 2017). A typical person from the European middle ages would have been absolutely certain of this plasma-like, spirit description of consciousness (Bailey, 2017). It is difficult to overstate the intuitive hold these ideas have had, and continue to have, on people. Over the past several hundred years, science has chipped away at these beliefs, leaving this social-cognitive model present at an implicit level but removing much of it from our intellectual belief structures. Most scientists and the educated public now understand properties 2 and 3 to be incorrect. Some people still believe them explicitly, but most of us have been taught otherwise. Most scientists and scholars, however, have yet to accept that property 1 may be just as much a component of a simplified, schematic model.

The reason for this scientific gap may be that properties 2 and 3, a mind that can flow through space and a mind that can transfer energy, are more easily experimentally tested. One can test for, and fail to find, mental auras and forces flowing outside the body—as, for example, Benjamin Franklin did when putting Mesmerism to the test (Franklin et al., 2002; Kihlstrom, 2002), or as Titchner and Coover did when putting the power of someone else's stare to the test (Coover, 1913; Titchner, 1898). Properties 2 and 3 are demonstrably false. Property 1, a magic experience that takes hold of information, does not refer to any specific, measurable effect on an outside object. It is therefore difficult (or impossible) to test experimentally. One cannot put it to a scientific debunking test. Hence, science is now in an awkward, inconsistent state. Or, at least, we see it partway through a long historical transition, with one foot still in the middle ages. Some aspects of the intrinsic model of mind are now scientifically dismissed as a ghost theory, while other aspects are still assumed to be literally true. Scholars are engaged in an effort to find the scientific basis for property 1: how does the brain generate a subjective experience that, itself, has no physical attributes? How does the brain generate qualia, or awareness, or consciousness? It may be time to recognize that the hard problem of consciousness belongs fundamentally to the same category as auras, mind beams, soul, ka, chi, and spirit.

We suggest that the almost universal approach of assuming that property 1 is true—taking it as axiomatic that a mind has a non-materialistic subjective experience and then trying to understand the mechanism behind it—is just as incorrect as assuming the validity of properties 2 and 3. It is like asking: how does the brain generate beams of mental energy that stream out of the eyes and affect other people? The assumption is wrong. The answer is that the brain constructs simplified models of its world and of itself. When we make claims about ourselves based on introspection, the brain is accessing and relying on those imperfect internal models. We must stop assuming that an introspected property is literally accurate; all we know is that the brain has constructed the information on which the claim is based, and the information is likely to be a simplified representation of something else. In the present argument, m-consciousness, the mysterious extra essence inside us, does not exist as such. Or at least, it is not what we think it is. We think we have it because of the self-descriptive models that the brain builds. I-consciousness is what the brain actually has; m-consciousness is what the brain thinks that it has.

This view resembles the illusionist view of consciousness (Dennett, 1991; Frankish, 2016). We are not saying, however, that consciousness is an illusion in the sense of something that does not exist. We are suggesting instead that i-consciousness exists, and that m-consciousness is the brain's natural, built-in, but imperfect understanding—an efficient and therefore detail-poor understanding—of i-consciousness. There is indeed someone home—but the someone is slightly misled about his or her exact nature.

We are also not saying that, because m-consciousness is an inaccurate version of reality, it should be

dismissed or ignored by science. Many scientists belittle the ghost intuitions, the mystery, and the hard problem, but we fundamentally disagree with that dismissive attitude because it ignores a major part of the scientific puzzle. To understand how the brain makes predictions about itself and other brain-controlled agents, we need to acknowledge that it constructs models, even while we acknowledge the inaccuracy and schematic nature of those models. To give an analogy, a map depicts a cartoonish distortion of a city. The city is not literally two-dimensional or composed of coloured lines, and yet the map is still useful. If you throw out the map because of its scientific inaccuracies, you might end up lost. We argue that m-consciousness stands to i-consciousness as the non-existent, simplified city stands to the actual city. It is the brain's quick-and-dirty, but useful model of i-consciousness.

The approach we are describing here resonates with many other people's approaches to consciousness. There is now a growing scientific convergence around a similar set of ideas, in which m-consciousness does not exist as such. Gazzaniga (1970) referred to an interpreter that constructs a self-narrative. Nisbett and Wilson (1977) suggested that everything we say about ourselves depends on partial and often incorrect information available to speech systems. Rosenthal (1991, 2005), Gennaro (2012), and many others suggested a higher-order thought theory of consciousness, in which we claim to be conscious because we contain meta-information about how we process information. Carruthers (2012) proposed a related, meta-cognition approach. Dennett (1991) wrote about consciousness as an illusion and the brain as a machine that falsely believes it contains subjective qualia. Frankish (2016) helped marshal scholars around illusionism as a theory of consciousness. Blackmore (2003) emphasized how memes, or ideas that accumulate culturally, may have given us the belief that we have conscious minds. Holland and Goodman (2003), from an engineering and robotics perspective, proposed that consciousness may be the presence of internal models used for prediction and control. Metzinger (2009) proposed that consciousness is related to internal models of arousal, and that the brain, captive to its own internal information, thinks it is conscious. Prinz (2017) and Frith (2002) emphasized social models of other people's minds as a basis for attributing consciousness to oneself. Chalmers (2018) put a useful, clarifying label to many of these approaches by posing the meta-problem, the question of why people believe we have a hard problem.

Most of these perspectives, and probably many others, share basic features. In our view, the most important feature is the following line of argument. Logically, the brain cannot put out a claim unless it contains the information on which the claim is based. Therefore, everything we think we know about ourselves, no matter how fervently we believe it, derives from internal information. Our certainty that we have a subjective experience, or qualia, or a "what it feels like", derives from internal information. That information, however, is not necessarily literally accurate. It is probably not, because the brain's models evolved to be efficient rather than accurate. In this account, people claim to have m-consciousness because we are information-processing machines that have constructed a schematic model of our i-consciousness. We are i-conscious of having m-consciousness, and m-consciousness is a model of i-consciousness.

Why a model of attention?

AST focuses on how the brain constructs a schematic model of attention (Graziano, 2013, 2019a, 2019b; Graziano & Kastner, 2011; Webb & Graziano, 2015). But why limit the theory to attention? The brain contains other processes such as decision-making, memory, and movement coordination. The brain might construct models of these other cognitive processes as well. Why not build a theory of consciousness around a decision-making schema, the brain's quick-and-dirty model of how it makes decisions? Or a memory schema, the brain's imperfect model of how it stores and retrieves memory? Here we will give several related answers that get at the heart of AST.

First, it is likely that the brain *does* build models of these other internal processes. However, they do not seem to correlate tightly with consciousness. At any particular moment, you may be engaged in cognitive decision-making or instead more passively experiencing the world around you, but in either case you can be subjectively conscious of something. You may be engaged in recalling a past memory or instead focused on events unfolding around you in real time, but again, in either case, you can be

equally subjectively conscious. Neither decision-making nor memory are tightly correlated with consciousness. In contrast, attention and consciousness have a closer relationship. As we will discuss below, the evidence suggests that attention and subjective awareness are tightly linked and difficult to separate. If you are directing attention toward something, you are likely to be conscious of it. If you are directing no attention toward something, you are unlikely to be conscious of it. You may think you are continuously aware of the full world around you regardless of how your attention is deployed, but that is not so. There are many now-classic experiments on what is called inattentional blindness, in which withdrawal of attention from an item leads to a loss of awareness of the item (e.g., Drew, Võ, & Wolfe, 2013; Mack & Rock, 2000; Simons & Chabris, 1999).

Many studies over the past several decades, including from our own lab, show that attention and subjective awareness can be dissociated (e.g., Ansorge & Heumann, 2006; Hsieh, Colas, & Kanwisher, 2011; Jiang, Costello, Fang, Huang, & He, 2006; Kentridge, Heywood, & Weiskrantz, 1999, 2004; Lambert, Naikar, McLachlan, & Aitken, 1999; McCormick, 1997; Tsushima, Sasaki, & Watanabe, 2006; Webb, Kean, & Graziano, 2016b). Attention is possible without awareness (although, thus far, there is no clear evidence of awareness without attention). People can direct attention to a visual stimulus, in the sense of focusing processing resources on it at the expense of processing other stimuli, while reporting a lack of subjective awareness of the stimulus. It is a mistake, however, to conclude that attention and awareness are independent. They are extremely difficult to separate. To hit that narrow window where the stimulus is strong enough to affect attention but not strong enough to trigger awareness, one must typically use stimuli that are masked or faint, titrated at the edge of detection. It is easier to separate the arm from the arm schema (Botvinick & Cohen, 1998; Graziano & Botvinick, 2002; Lackner, 1988) than it is to separate attention from awareness. In AST, the reason why awareness usually (but not always) tracks attention is that awareness is a construct, a model, whose purpose is to represent attention. Only when the model makes a mistake does attention occur without awareness.

Another way to get at the relationship between consciousness and attention is to consider the properties of attention in more detail. Attention, as a mechanistic process in the brain, is not monolithic. Psychologists have studied covert and overt attention, spatial and feature attention, exogenous and endogenous attention, visual, auditory, and tactile attention, and attention to other, more cognitive, internal events such as decisions and plans (Norbre & Kaster, 2014). If the brain is to model itself, why would it construct a single attention schema, instead of a large collection of models representing different kinds and aspects of attention? In AST, one general aspect of attention is of greatest behavioural importance. Attention, at the cortical level, is a competition among chunks of information (Beck & Kastner, 2009; Desimone & Duncan, 1995). That competition occurs over many layers, across complex cortical hierarchies. At any moment, one or a few sets of information have won the global competition, are subject to deep cortical processing, and are able to maximally impact decision-making, memory, and behaviour. In AST, it is that general, most behaviorally-relevant aspect of attention that is modelled. The brain has no need for a scientifically detailed model of attention that differentiates the many types and mechanisms. Instead, attention is modelled inaccurately, but efficiently, as a single, diffuse thing that substantially drives behaviour.

Attention, at that most general level, has a set of properties closely matching the properties typically attributed to subjective consciousness.

(1) Both attention and awareness are directed at a target. A person attends *to* something, and is aware *of* something.
(2) Both attention and awareness are products of an agent. Attention is an emergent computational property of the brain. Awareness, as most people intuitively understand it, implies an "I" who is aware.
(3) Both attention and awareness are selective. Only a tiny fraction of the large amount of information available at any one time is attended. Awareness is selective in the same way. One is aware of only a small amount of the information flowing into the senses or generated internally.
(4) Both attention and awareness are graded. Attention typically has a single focus, but can be distributed to some extent to secondary targets. Awareness also has a focus and is graded. A

person can be most intently aware of A and a little aware of B.

(5) Both attention and awareness operate over the same information domains. Beyond vision, one can attend to stimuli in any of the five senses, to a thought, to an emotion, or to a recalled memory, to give a few examples. One can also be aware within the same information domains.

(6) Both attention and awareness impact decision-making and behaviour. When the brain attends to something, the neural signals are enhanced, gain greater influence over the down-stream circuitry, and have a greater impact on behaviour. Likewise, in the common intuitive understanding of awareness, when you are aware of something, you can choose to act on it. In both cases, the implication is that they are fundamental driving forces in behaviour.

(7) Both attention and awareness imply deep processing. Attention is when the brain devotes more computing resources to a selected information set, arriving at a deeper analysis. Awareness implies a mind seizing on something in the sense of being occupied by it, or vividly experiencing it, or taking possession of it.

And yet despite all the similarities, attention and awareness are not the same thing. A crucial difference between them is that attention is a physical, objectively measurable process in the brain, whereas awareness is something we access introspectively and say that we have. Logically, when people claim that they have a subjective experience, that claim must be based on information constructed within the brain, or they wouldn't be able to make the claim. That information is effectively a type of self model. It is information descriptive of the self. But what actual, physical component of the self is being modelled? The answer is: whatever physical component of the self most closely correlates with that information. Hence we arrive once again at the same hypothesis: awareness is a model of attention, because attention is the physical process that most closely correlates with the claim of awareness. Awareness is as similar to attention as, for example, the arm schema is to a physical arm. One is a construct that represents the other, tracking it closely most of the time. The central proposal of AST is that the brain not only uses attention, but also constructs a general schematic model of it, and that model supplies the information on the basis of which we believe we have a subjective experience.

Why are we so certain it's real?

Any mechanistic theory of consciousness inevitably encounters a fundamental philosophical challenge. To most people, m-consciousness is not something we *think* we have, in a cognitive manner, but something we *do* have. No amount of cultural learning or voluntary thought can erase it. For example, one cannot intellectualize away the sensation of pain. This common certainty that we actually do have m-consciousness begs the question: why are people so certain that we have it? Never mind for the moment whether we actually do have it. In AST, in a sense we do and in a sense we don't. We have i-consciousness, and m-consciousness is a distorted picture of i-consciousness. But regardless of the actual presence or absence of it, why are we certain we have it? When you look at something red, why is it so compellingly obvious that you have a subjective experience of red?

Consider the cognitive differences between something that people consider to be "real" and something that people consider to be merely an internal construct. For example, when you look at an apple, you do not typically think, "I have intellectually thought up the sight of the apple". The low-level model of the apple, constructed in the visual system, may be associated with at least two properties that label it as reality rather than self-construct. First, one cannot turn it off. The representation is not cognitively modifiable. You cannot voluntarily change a visual model of an apple into a butterfly by cognitive fiat. The second property relates to source monitoring, or reality monitoring, a process that has been extensively studied in psychology (Simons, Garrison, & Johnson, 2017). The brain not only constructs representations, but also constructs information about the possible source of a representation. Source monitoring allows us to distinguish between the hypothetical and the real. It is the largely hidden process that allows you to say, "I'm looking at a real apple", versus, "An apple construct is in my mind—I thought it up". Intriguingly, source monitoring can make spectacular errors, such as when people are certain of the reality of false memories (Simons et al., 2017). Likewise, people who suffer from a sense of unreality of the world around them, or

on the opposite end, people who are prone to a type of hallucination in which imagination is taken to be real, may be suffering from an error in source monitoring. Just because you are certain something is real or unreal does not make it so—it means only that your source monitoring mechanism has settled on one particular answer.

We suggest that the reason most people consider m-consciousness to be something they actually have is the same reason why people are certain of the reality of anything. First, the mechanism that constructs the m-consciousness self-model is cognitively impenetrable. One cannot voluntarily turn it off or intellectualize it away. It is automatically present. Second, that model is subject to the normal process of source monitoring. The brain represents it as "real".

A common argument in favour of the reality of m-consciousness could be put as follows: "I know I have an experience because, Dude, I'm experiencing it right now". Every argument in favour of the literal reality of subjective experience, that the authors of this article have ever encountered, boils down sooner or later to that logic. But the logic is circular. It is literally, "X is true because X is true". If that is not a machine stuck in a logic loop, we don't know what is. The machine accesses internal information, the information describes a simplified, ghost version of i-consciousness, and the machine reports that version. Introspection (cognition accessing internal information) can never return any other answer, because the brain is captive to the information contained within it. On introspection, one will always arrive at that certainty.

Control of attention

What might be the adaptive advantage for the brain to construct an attention schema? Much of the experimental work in our lab is focused on this question of the cognitive role of an attention schema. We suggested two general possible cognitive uses: a better control of one's own attention through predictive modelling, and a better theory of mind through modelling the attentional processes of others.

A fundamental principle of control theory is that a controller works better if it incorporates an internal model (Camacho & Bordons Alba, 2004; Conant & Ashby, 1970; Francis & Wonham, 1976). A self-driving car works better if the controller contains a set of information that represents, or models, the dynamics of the car and how it interacts with the environment. The brain can better control movement because it constructs an internal model of the body and how the body interacts with the physical world (Graziano & Botvinick, 2002; Head & Holmes, 1911; Holmes & Spence, 2004; Shadmehr & Mussa-Ivaldi, 1994). Attention is a highly complex, dynamic process that is directed by control systems within the brain. It is exactly the type of control problem that could benefit from an internal model. We proposed (Graziano, 2013, 2016; Graziano & Kastner, 2011; Webb & Graziano, 2015) that the brain constructs 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. We labelled this proposed internal model the "attention schema" in analogy to the body schema that is theorized to contribute to the control of movement. In order to explain this analogy, below we provide a summary of the body schema and some of the ways it has been studied.

Many lines of evidence suggest the existence of the body schema. One obvious source of evidence is that people can report the state of a limb. Close your eyes and you can easily report on the size, shape, and specific position of your arms. The ability to report the state of one's arm is not merely dependent on a stream of sensory information, as can be seen particularly clearly in the case of a phantom limb. When the limb is amputated, many patients can report on the state of a limb that, apparently, exists only in the form of an internal model (Medina & Coslett, 2016; Ramachandran & Rogers-Ramachandran, 2000). The same process can occur in reverse—patients with damage to the parietal lobe can sometimes suffer from somatoparaphrenia, a condition in which a limb is physically present and low-level sensory input is intact, but the patient feels no ownership of the limb, possibly reflecting damage to the internal schema of the limb (Vallar & Ronchi, 2009). The body schema has also been studied in the non-clinical population. Illusions of the body, such as the rubber hand illusion (Botvinick & Cohen, 1998) and the Pinocchio illusion (Lackner, 1988), suggest that the body schema relies on a sophisticated integration of information from vision, touch, proprioception, motor feedback, and learned, stable properties about the

shape and structure of a limb. Some aspects of the body schema may be especially useful for the predictive avoidance of collisions with nearby objects (de Vignemont, 2018).

The brain can alter its model of the body through experience. This adaptation has been especially thoroughly studied with respect to the arm (Mazzoni & Krakauer, 2006; Shadmehr & Mussa-Ivaldi, 1994; Taylor, Krakauer, & Ivry, 2014; Thoroughman & Shadmehr, 2000; Thoroughman & Taylor, 2005). If a person reaches toward a target, the reach is typically fast, straight, and accurate. But if the target is displaced by prism glasses, or if the reach is deviated by a subtle force field applied to the arm, the person will mis-reach. Very quickly, within a few trials, the person will adapt and reach accurately again. A standard interpretation of this result is that the motor system contains a model of how the arm functions during reaching. When that model is no longer accurate and errors in reaching occur, the model is then modified through learning. Arguably one of the most important functional advantages of an internal model is that it allows the control system to adapt easily to changing circumstances.

This brief summary of the arm model literature illustrates the range of content and function relevant to an internal model. Studies of the arm model have included adaptation, control of reaching, perception of limb size and shape, and collision avoidance. Some components are explicitly verbalizable by participants, and some are implicit. All of these components together form a complicated, rich representation. Much remains to be learned about the body schema. Despite the substantial unknowns, however, the realization that the brain constructs an internal model of the body, often credited to Head and Holmes in 1911, has been a crucial organizing principle in studying the motor system for more than 100 years. We suggest that a similar insight can be applied to the control of attention. In that suggestion, the brain constructs an attention schema to help in the control of attention. Like the body schema, the attention schema is constructed automatically. We do not have cognitive control over it. Also like the body schema, at least some aspects of the proposed attention schema are cognitively accessible and verbalizable. Because of the imperfect manner in which the attention schema depicts attention, people claim to have a vague, non-physical essence of subjective experience inside of them, instead of accurately describing the mechanism of attention.

If AST is correct, then the relationship between subjective awareness and attention should be similar to the relationship between the arm schema and the arm. Although the two usually correlate, the arm schema can be dissociated from the arm. When the arm schema fails, the control of the arm is still possible, but is compromised. Awareness should usually track attention, but under some conditions the two should be dissociable. Without awareness of a stimulus, attention to that stimulus should still be possible, but the endogenous control of that attention should be compromised.

As noted in the previous section, it is now well established that attention to a stimulus can occur without subjective awareness of the stimulus (e.g., Ansorge & Heumann, 2006; Hsieh et al., 2011; Jiang et al., 2006; Kentridge et al., 1999, 2004; Lambert et al., 1999; McCormick, 1997; Tsushima et al., 2006; Webb et al., 2016b). However, the evidence does not point to a simple independence of attention and awareness. At least some studies suggest that the two interact. The reported interactions are, in our interpretation, consistent with AST, because without awareness of a stimulus, although attention to the stimulus can remain, the control of attention toward that stimulus is compromised.

An especially informative example involves attention to a distractor stimulus. One of the most frequent challenges the attention system faces is to reduce attention to a distractor. Think of diverting small amounts of attention to minor distractions around the room, such as a dog and a mosquito, while focusing attention mainly on the central task at hand, an important phone conversation. One study suggests that when people are unaware of the distractor stimulus, more of their attention leaks toward the distractor and away from the task-relevant stimuli (Tsushima et al., 2006). At first the finding may seem counterintuitive. When people are *aware* of the stimulus, they direct *less* attention to it. When they are *unaware* of it, they direct *more* attention to it. Shouldn't awareness increase attention? The finding, however, perfectly aligns with the control-theory interpretation. If awareness acts as the brain's model of attention, then being unaware of the distractor stimulus means there is a gap in the model—a failure to model attention to the distractor. That gap

in the model leads to a poor ability to regulate and minimize attention to that distractor.

In another study, without awareness of a stimulus, attention to the stimulus was not overall smaller or larger in magnitude, but showed greater fluctuations over time, possibly reflecting a reduction in control and stabilization of attention (Webb et al., 2016b).

Studies like these convincingly show several properties of attention and awareness. First, the two are not the same thing. They can be dissociated. Second, awareness is not simply the upper end of attention, or an especially enhanced state of attention. Sometimes the presence of awareness can even decrease the degree of attention to a stimulus. Third, awareness has a substantial effect on the control of attention. Without awareness of a stimulus, attention to that stimulus is no longer controlled as well for the needs of the ongoing task. That relationship is consistent with the hypothesis that awareness serves as the control model for attention. The findings, of course, do not prove AST. Many alternative explanations of the same results may be possible. However, AST has the advantage of making a simple, underlying sense of the otherwise complex relationship between awareness and attention.

Modelling the attention of others

We proposed that an attention schema would be useful in modelling the attentional state of others and thus predicting the behaviour of others (Graziano, 2013, 2019b; Graziano & Kastner, 2011). To draw an analogy to the body schema once again, it is worth noting that the body schema is involved in a relatively little-known social phenomenon. When judging the postures of other people's bodies, we recruit the same brain mechanisms that construct our own body schema (Bonda, Petrides, Frey, & Evans, 1995; Parsons, 1987; Sekiyama, 1982). In the case of the body schema, modelling others appears to be a minor extension of the mechanism for modelling oneself. In the case of the attention schema, we argued that modelling others is a much more prominent extension to the mechanism, and is a major part of the social toolkit (Graziano, 2013). In that suggestion, people not only attribute m-consciousness to themselves, but also attribute it to others. We live in a sea of perceived consciousness that we paint onto ourselves, others, and sometimes even inanimate objects and empty spaces that are the targets of an exuberant social cognitive process. We have a hair trigger for attributing consciousness, because it is so socially useful that it is better to mistakenly overuse it than mistakenly underuse it.

Attributing consciousness to others and to oneself, however, are obviously not identical processes. Self-attribution has more layers due to its closed-loop nature. An attention schema directed at the self is useful not only to predict, but also to control oneself. Moreover, a richer source of information is available to construct one's own attention schema, beyond the simple visual cues that we can register from other people. These considerations suggest that the consciousness we attribute to others is likely to be a pale version of the consciousness we attribute to ourselves.

Modelling the attention of others is one component of theory of mind, the ability to attribute beliefs, intentions, emotions, goals and agendas to others (Baron-Cohen, 1997; Frith & Frith, 2003; Premack & Woodruff, 1978; Wellman, 2018; Wimmer & Perner, 1983). How people reconstruct the attention of others, however, is often studied in a limited manner, for example treating the direction of gaze as a proxy for visual attention (e.g., Baron-Cohen, 1997; Calder et al., 2002; Friesen & Kingstone, 1998; Frischen, Bayliss, & Tipper, 2007). But modelling someone else's attention is a far more complex and rich process than tracking gaze direction. Gaze direction is just a cue that can be used to help constrain a model of attention.

To convey something of the richness of the task of modelling attention, suppose I am watching a person in a coffee house with a doughnut on the table in front of him. His eyes are on the doughnut. Is his attention really on the snack, or is it covertly on his ex-girlfriend who just walked in the door? Or is his attention on a thought or a memory, unrelated to his gaze direction? Is his attention now suddenly drawn exogenously to someone waving from across the room? Is his attention directed by his own choice, endogenously, as he searches the room for a friend? If I think he's finally about to reach for that doughnut, and then his cell phone suddenly rings in his pocket, can I presume his attention is suddenly pulled to the phone and away from the pastry, thereby reducing the chance that he'll reach at that exact moment? Once his attention is attracted to the phone, even if he chooses not to answer it, can I intuitively guess that his attention has some viscosity and will linger on the phone for at

least a good half second, before he is likely to re-direct attention back to the doughnut? Do I understand that his attention can be focused or divided, but is a limited resource, almost like a fluid of limited volume flowing out of him that can be spread thinly or focused intently? Do I understand that in the most general sense, his attention is his mind taking possession of items, thereby enabling him to make choices and react? Do I understand that attention does not always lead to an immediate action, but leads to the ability to choose, and can lead to information being stored in memory that might drive future behavioural choices? All of these pieces together constitute a dynamic model of attention. That model can be constrained by details like where his eyes are directed, or his facial expression, but the model itself is a rich, largely implicit understanding of how his mind takes possession of items. The model of attention is not a model of his emotions, or his decisions, or his intentions, or his knowledge of the world—all the content that is often associated with theory of mind. The model of attention is in some sense more fundamental. It precedes these more specific components. It is a model of what it means for him to have a mind that can contain any content at all.

Several recent experiments suggest that people do indeed construct a model of the attention of others. The model is constrained by specific visual details, and the model can contain some quirky, physically unrealistic attributes (Guterstam, Kean, Webb, Kean, & Graziano, 2018; Kelly, Webb, Meier, Arcaro, & Graziano, 2014; Pesquita, Chapman, & Enns, 2016). One study (Pesquita et al., 2016) found that when subjects watched a video of an actor attending to an object, the subjects implicitly encoded whether the actor's attention was drawn to the object exogenously (by the salience of the object) or was directed endogenously (by the actor's own choice). Here people were constructing an implicit model of an agent's attention —not just information about the object of attention or the direction of gaze, but information about the process of attention itself.

A recent study of ours (Guterstam, Kean, Webb, Kean, & Graziano, 2018) found that when people viewed a face gazing attentively at an object, they treated the stimulus configuration as though an invisible, gentle, mind-force were emanating from the face and physically pushing on the object. That perceived force was revealed when subjects were asked to make physical judgments about the object and how it might tip over. When quizzed explicitly, the subjects showed no knowledge that they were treating the attentive face as though it were radiating a beam at the object. Yet the results were consistent with subjects constructing that implicit model of attention. When people were told that the face was not attending to the object, but instead at a different, more distant object, the effect disappeared. Likewise, if the face in the display was blindfolded or turned away, the effect was not observed. The findings suggest that people construct a descriptive model of the visual attention of others, the model is constructed implicitly and automatically, and the model contains some physically incorrect and schematic features. Other agents are represented as a source of an energy-like essence associated with attention, that radiates invisibly through space where the agent directs it, and that touches and even physically affects the object of attention. As bizarre as this implicit perception may seem to modern sensibilities, it is consistent with thousands of years of intuitions and assumptions about how the mind works.

We suggest that this model of other people's attention, as a fluid-like substance that is generated inside of an agent and flows out toward targets, may be a useful simplification, a geometric trick for keeping track of who is attending to what and by how much. It is certainly an easier model to construct on the fly than a scientifically accurate account of attention as a matrix of billions of neuronal and synaptic interactions. It would do the human species no good to have evolved a model of attention that is neuroscientifically accurate, whereas it may well be pragmatically useful to know intuitively that person A has a "beam" of attention directed at object B. This "fluid flow" model of attention, as inaccurate as it is, may be a useful and efficient trick for keeping track of who is attending to what in a complex social environment. The data thus far tend to support the idea that people construct a simplified model of the attention of others, much of the model is constructed at an implicit level, and at least some aspects of the model are schematic and extremely physically inaccurate. Even though this model is automatic and partly implicit, we suggest it may have played a significant role in biasing intuitions and therefore in shaping culturally common ideas about mind and consciousness. The belief in a hard problem of consciousness, we

suggest, owes itself partly to the deeply engrained intuition that the mind is a physically ghost-like, invisible essence that is generated inside of an agent.

Where and how is the attention schema constructed?

What we mean here by a model is obviously not a literal picture in the brain, but a set of information contained in a neural network. Imagine building a deep-learning neural network—call it network A—that engages in artificial visual attention. It receives visual signals, the signals compete through many internal layers, and some signals reach such a state of enhancement that they dominate the network, something like when information in the brain enters the global workspace. Although we are describing this hypothetical network somewhat vaguely here, artificial attention systems have been constructed before many times (e.g., Borji & Itti, 2013; Deco & Rolls, 2004; Le Meur, Le Callet, & Barba, 2006; Reynolds & Heeger, 2009; Schwemmer, Feng, Holmes, Gottlieb, & Cohen, 2015).

Now imagine a second neural network—call it network B—whose job is to make predictions about the attentional dynamics of network A. Crucially, the job of network B is not to re-describe the visual information that percolates through network A. It is not a higher-order, re-representation of visual stimuli. Instead, network B builds a set of information descriptive of the process of attention itself. It is used to feed back on and help control the attention process in network A. Here again, such an artificial system is implementable. For example, one recent model includes artificial attention and an internal model of attention that is used to help regulate the control of attention (van den Boogaard, Treur, & Turpijn, 2017).

For user convenience, such that we can "talk" to our hypothetical construction, consider a third network, network C, whose job is to receive output signals from both network A and B and transform them into a user-friendly format such as speech. Now we have a complex system with three components, each of which has many layers. Network A is something like the visual system, which contains visual information and uses visual attention. Network B constructs an attention schema—a predictive model of how network A deploys attention. Network C allows the system to report to the outside world. However, the system can only report the information contained within it. It has no accurate, scientifically precise information on the visual world or on its own attention. It can, at best, report schematically encoded visual properties, and it can report whatever distorted, schematic, or efficient information it has constructed on its own attention processes. This is the kind of architecture we are proposing for AST. The machine does not experience anything. It has no qualia. It has no phenomenology. But it is busy with internal information—it *thinks*. It thinks there are visual objects surrounding it, because that is what its internal information indicates. It thinks it has subjective experiences of those objects, because, again, that is what its internal information indicates. It thinks—and claims—that it has m-consciousness.

Where in the brain might we look for the crucial network B, the network that constructs an attention schema? We previously suggested that a cortical network overlapping part of the temporoparietal junction (TPJ) may contribute to the computation (Graziano & Kastner, 2011; Kelly et al., 2014; Webb, Igelström, Schurger, & Graziano, 2016a). This suggestion was based on a confluence of three properties. First, in AST, an attention schema helps in the control of one's own attention. Second, it helps in social cognition by modelling the attention of others. Third, it is responsible for one's construct of subjective awareness. Thus a network that constructs an attention schema might be involved in one's own attention, in social cognition, and when it is damaged, one might expect a disruption of awareness. The TPJ combines all three properties. Cortical networks involved in theory of mind pass through the TPJ (Molenberghs, Johnson, Henry, & Mattingley, 2016; Saxe & Kanwisher, 2003; Saxe & Powell, 2006). Networks involved in attention, especially the ventral attention network, the saliency network, and the control network, pass through the TPJ, and have at least some overlap or interaction with regions involved in social cognition (Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000; Igelström & Graziano, 2017; Igelström, Webb, Kelly, & Graziano, 2016; Shulman et al., 2010). Damage to the right TPJ can lead to the longest lasting and most severe cases of hemispatial neglect, which is arguably the most severe clinical disruption of a person's awareness (Vallar & Perani, 1986). Brain imaging studies from our own lab suggest that when people attribute

awareness to others, a cortical network is recruited that passes through the TPJ (Kelly et al., 2014), and when visual awareness is manipulated, similar regions of the TPJ are implicated (Webb et al., 2016a). For these reasons we tentatively suggested that an attention schema may be constructed primarily in a cortical network that includes regions within the TPJ and probably connected regions of the prefrontal cortex. However, that anatomical proposal remains a first-pass hypothesis. We recognize that it is inadvisable to pin a specific function on a brain network without more data.

Finding common ground

For the past ten years we have argued in favour of AST as an explanation of consciousness. However, we are not arguing against all other theories of consciousness. A subset of prominent theories may fit well together. We find a deep commonality between HOT, GW, AST, and the illusionist approach to consciousness. Here we suggest that these theories can be understood as different, interlocking perspectives on the same underlying mechanism.

In GW (Baars, 1988; Dehaene, 2014; Dehaene & Changeux, 2011), information is boosted and stabilized by exogenous or endogenous attention mechanisms until it reaches the global workspace, where it becomes available to many systems including speech, decision-making, movement control, and memory. In this perspective, information that has entered the global workspace has entered subjective awareness. Awareness corresponds to the highest level of attention in the brain, in which information has been enhanced to a threshold level, sometimes called ignition.

One of the weaknesses of GW, at least in its simplest form, is that it leaves unexplained how people end up believing they have a subjective experience. GW is analogous to the network A from the previous section. GW accommodates how the brain can focus resources on a stimulus, process it deeply, and report the properties of that stimulus. But there is no simple explanation for why we claim to have an added subjective experience of the stimulus. GW explains i-consciousness while lacking an obvious relationship to m-consciousness. AST supplies that extra piece. By positing an added network B that constructs an attention schema, AST explains how the brain can think it contains m-consciousness—the mysterious, physically impalpable essence attached to most items that receive a measure of attention.

A second prominent theory of awareness, HOT (Gennaro, 2012; Lau & Rosenthal, 2011; Rosenthal, 1991, 2005), depends on the insight that the claim, "I am aware of the stimulus", contains more information than the claim, "There is a stimulus". In the theory, the claim of awareness requires higher-order information about one's own internal processes in addition to lower-order information about the stimulus. Awareness derives from that higher-order representation. Exactly what that higher-order information is, what cognitive purpose it serves, and where in the brain it may be constructed, is in debate, though some associate the higher-order processes with the prefrontal cortex (Lau & Rosenthal, 2011; Odegaard, Knight, & Lau, 2017).

When AST was first proposed, it was described from the perspective of social cognition (Graziano, 2013; Graziano & Kastner, 2011). Abel attributes the property of awareness to Bill as a simplified, but useful model of Bill's attention. Just so, Abel attributes the property of awareness to himself as a simplified, useful model of his own attention. A model of someone else's attention would be useful for making predictions about that other person's behaviour. A model of one's own attention would be useful for predicting one's own behaviour, and would also be useful for regulating attention itself—much like the body schema is fundamental to controlling movement of the limbs.

These three theories seem superficially to have little overlap. And yet, at a deeper level, they connect. There is nothing complicated about that connection. AST can be understood as a specific unification of GW and HOT. In AST, the brain contains attention, which boosts signals ultimately into a global workspace. However, in addition, the brain also constructs a higher-order representation of that global workspace. That higher-order representation is not a representation of the specific contents inside the global workspace; it is not, for example, a re-representation of the apple you are looking at. Instead, it is a representation of the dynamics and consequences of having a global workspace. AST is literally the simplest possible way to unify HOT and GW, in that it posits a higher-order representation of the global workspace. The global workspace is i-consciousness. We think we have a more mysterious, physically impalpable m-consciousness

because of the higher-order representation of the global workspace.

In this account, then, how does a person become visually conscious of an apple he is looking at? Suppose information about the apple, processed in the visual system, reaches the highest levels of attentional enhancement, and thus reaches the global workspace. There are at least a dozen cortical networks that span the parietal and frontal lobes (Bzdok et al., 2013; Igelström & Graziano, 2017; Yeo et al., 2011), but the networks most often proposed to be associated with the global workspace tend to overlap those involved in attention, including the dorsal attention network, the parieto-frontal executive network, the salience network, and the ventral attention network (Dehaene, 2014; Dehaene & Changeux, 2011; van Vugt et al., 2018). Some researchers emphasize the prefrontal component of these networks, especially the dorsolateral prefrontal cortex (Lau & Rosenthal, 2011; Odegaard et al., 2017; van Vugt et al., 2018). The apple information, reaching these networks, has reached a central and prominent place where it can impact output systems, giving the person the ability to talk about the apple, to reach for it, and to make high-level cognitive decisions about it. Thus far, we have an explanation for i-consciousness.

But when the person claims that there is an extra essence, a what-it-feels-like, the qualia associated with the apple's colour and shape, something non-materialistic—m-consciousness—that added claim should, in the present proposal, require information constructed within a specific network. In our current hypothesis, that computation is carried out by the theory-of-mind network, which includes such brain areas as the temporoparietal junction, the superior temporal sulcus, and the dorsomedial prefrontal cortex (Igelström & Graziano, 2017; Molenberghs et al., 2016; Saxe & Kanwisher, 2003; Saxe & Powell, 2006). In that hypothesis, the very notion of a conscious mind, and the human claim to have one, depends on the theory-of-mind network.

Neuroscientists who study consciousness traditionally ask a specific question: where in the brain must information be sent, to generate subjective experience? If I am aware of an apple, is the experience generated when the apple information is processed in low-level sensory areas, or when it reaches higher-level cortical areas associated with cognition—or, perhaps, when it enters some other anatomical substrate in the brain? When asking that traditional question, it makes little sense to propose that the theory-of-mind network gives us consciousness. Why would the apple information ever enter the theory-of-mind network, and even if it did get there, why would it generate consciousness? In the present proposal, however, the traditional question is ill-posed. No matter what brain area or network receives the information about the apple, it never generates a subjective feeling of the apple. Instead, a distinct brain system computes a specific information set about what subjective experience is. M-consciousness becomes another computed property ultimately linked to the apple, like colour or motion or spatial location. In the present account, without the theory-of-mind network building a model of what a conscious mind is and adding that information to the global mix, the person would have no basis to make any claims about having a conscious experience of the apple.

In this perspective, the global workspace could be considered more like a soup pot. Ingredients may be pre-cooked in other pots, but eventually some are combined in the central pot. For a person to claim to be conscious of the apple, that central pot must have received information about the apple *and* information about consciousness. The information about the apple, which is cooked in the visual system, serves as a model of a specific external object. The information about consciousness, which is cooked in the theory-of-mind network, serves as a model of the central pot itself. That model is so simplified and schematic that it describes the pot as a non-material essence, m-consciousness. Combined, the two models provide sufficient information to form the basis of the claim, "I have a subjective, conscious experience of that apple".

In this article, we emphasized AST, a theory we have been developing for some years. However, the theory should not be viewed in isolation. It has direct links to other prominent theories, including GW and HOT. It is also a type of illusionist theory because people do not actually *have* m-consciousness, but instead *attribute* m-consciousness to themselves. We do not view these many theories as rivals. They are more like different keyhole perspectives on a single, underlying mechanism. Contrary to the platitude that science does not yet understand consciousness, we suggest that a subset of theories and ideas already point

toward a core explanation. We may now have a "standard model" of consciousness—a family of theories that cohere and provide a working, mechanistic, scientifically meaningful, and even artificially buildable understanding of consciousness.

Acknowledgements

Supported by the Princeton Neuroscience Institute Innovation Fund.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Supported by the Princeton Neuroscience Institute Innovation Fund.

References

Ansorge, U., & Heumann, M. (2006). Shifts of visuospatial attention to invisible (metacontrast-masked) singletons: Clues from reaction times and event-related potentials. *Advances in Cognitive Psychology*, *2*, 61–76. doi:10.2478/v10053-008-0045-9

Baars, B. J. (1988). *A cognitive theory of consciousness*. New York: Cambridge University Press.

Baars, B. J., & Franklin, S. (2007). An architectural model of conscious and unconscious brain functions: Global workspace theory and IDA. *Neural Networks*, *20*, 955–961. doi:10.1016/j.neunet.2007.09.013

Bailey, M. D. (2017). *Fearful spirits, reasoned follies: The boundaries of superstition in late medieval Europe*. Cornell: Cornell University Press.

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. Cambridge, MA: MIT press.

Beck, D. M., & Kastner, S. (2009). Top-down and bottom-up mechanisms in biasing competition in the human brain. *Vision Research*, *49*, 1154–1165. doi:10.1016/j.visres.2008.07.012

Benassi, V. A., Sweeney, P. D., & Drevno, G. E. (1979). Mind over matter: Perceived success at psychokinesis. *Journal of Personality and Social Psychology*, *37*, 1377–1386. doi:10.1037/0022-3514.37.8.1377

Blackmore, S. J. (2003). Consciousness in meme machines. *Journal of Consciousness Studies*, *10*, 19–30.

Bonda, E., Petrides, M., Frey, S., & Evans, A. (1995). Neural correlates of mental transformations of the body-in-space. *Proceedings of the National Academy of Sciences, U. S. A.*, *92*, 11180–11184. doi:10.1073/pnas.92.24.11180

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*, 185–207. doi:10.1109/TPAMI.2012.89

Botvinick, M., & Cohen, J. (1998). Rubber hands 'feel' touch that eyes see. *Nature*, *391*, 756. doi:10.1038/35784

Bzdok, D., Langner, R., Schilbach, L., Jakobs, O., Roski, C., Caspers, S., … Eickhoff, S. B. (2013). Characterization of the temporo-parietal junction by combining data-driven parcellation, complementary connectivity analyses, and functional decoding. *Neuroimage*, *81*, 381–392. doi:10.1016/j.neuroimage.2013.05.046

Calder, A. J., Lawrence, A. D., Keane, J., Scott, S. K., Owen, A. M., Christoffels, I., & Young, A. W. (2002). Reading the mind from eye gaze. *Neuropsychologia*, *40*, 1129–1138. doi:10.1016/S0028-3932(02)00008-8

Camacho, E. F., & Bordons Alba, C. (2004). *Model predictive control*. New York, NY: Springer Publishing.

Carruthers, G. (2012). A metacognitive model of the sense of agency over thoughts. *Cognitive Neuropsychiatry*, *17*, 291–314. doi:10.1080/13546805.2011.627275

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, *2*, 200–219.

Chalmers, D. (2018). The meta-problem of consciousness. *Journal of Consciousness Studies*, *25*, 6–61.

Conant, R. C., & Ashby, W. R. (1970). Every good regulator of a system must be a model of that system. *International Journal of Systems Science*, *1*, 89–97. doi:10.1080/00207727008920220

Coover, J. E. (1913). "The feeling of being stared at": Experimental. *The American Journal of Psychology*, *24*, 570–575. doi:10.2307/1413454

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience*, *3*, 292–297. doi:10.1038/73009

Deco, G., & Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, *44*, 621–642. doi:10.1016/j.visres.2003.09.037

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

Dehaene, S., & Changeux, J.-P. (2011). Experimental and theoretical approaches to conscious processing. *Neuron*, *70*, 200–227. doi:10.1016/j.neuron.2011.03.018

Dennett, D. C. (1991). *Consciousness explained*. New York: Little-Brown.

Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. doi:10.1146/annurev.ne.18.030195.001205

de Vignemont, F. (2018). *Mind the body: An exploration of bodily self-awareness*. Oxford: Oxford University Press.

Drew, T., Võ, M. L., & Wolfe, J. M. (2013). The invisible gorilla strikes again: Sustained inattentional blindness in expert observers. *Psychological Science*, *24*, 1848–1853. doi:10.1177/0956797613479386

Dundes, A. (1981). *The evil eye: A folklore casebook*. New York: Garland Press.

Francis, B. A., & 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. *Journal of Consciousness Studies*, *23*, 1–39.

Franklin, B., Majault, M. J., Le Roy, J. B., Sallin, C. L., Bailly, J.-S., d'Arcet, J., … Lavoisier, A. (2002). Report of the commissioners charged by the King with the examination of animal magnetism. *International Journal of Clinical and Experimental Hypnosis*, *50*, 332–363. doi:10.1080/00207140208410109

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin Review*, *5*, 490–495. doi:10.3758/BF03208827

Frischen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: Visual attention, social cognition, and individual differences. *Psychological Bulletin*, *133*, 694–724. doi:10.1037/0033-2909.133.4.694

Frith, C. (2002). Attention to action and awareness of other minds. *Consciousness and Cognition*, *11*, 481–487. doi:10.1016/S1053-8100(02)00022-3

Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London B Biological Sciences*, *358*, 459–473. doi:10.1098/rstb.2002.1218

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

Gennaro, R. (2012). *The consciousness paradox: Consciousness, concepts, and higher-order thoughts*. Cambridge, MA: MIT Press.

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

Graziano, M. S. A. (2016). Consciousness engineered. *Journal of Consciousness Studies*, *23*, 98–115.

Graziano, M. S. A. (2019a). *Rethinking consciousness: A scientific theory of subjective experience*. New York: Norton.

Graziano, M. S. A. (2019b). Attributing awareness to others: The attention schema theory and its relationship to behavioral prediction. *Journal of Consciousness Studies*, *26*, 17–37.

Graziano, M. S. A., & Botvinick, M. M. (2002). How the brain represents the body: Insights from neurophysiology and psychology. In W. Prinz, & B. Hommel (Eds.), *Common mechanisms in perception and action: Attention and performance XIX* (pp. 136–157). Oxford: Oxford University Press.

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

Gross, C. G. (1999). The fire that comes from the eye. *The Neuroscientist*, *5*, 58–64. doi:10.1177/107385849900500108

Guterstam, A., Kean, H. H., Webb, T. W., Kean, F. S., & Graziano, M. S. A. (2018). Implicit model of other people's visual attention as an invisible, force-carrying beam projecting from the eyes. *Proceedings of the National Academy of Sciences, U. S. A.*, *116*, 328–333. doi:10.1073/pnas.1816581115

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

Holland, O., & Goodman, R. (2003). Robots with internal models: A route to machine consciousness? *Journal of Consciousness Studies*, *10*, 77–109.

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

Hsieh, P., Colas, J. T., & Kanwisher, N. (2011). Unconscious pop-out: Attentional capture by unseen feature singletons only when top-down attention is available. *Psychological Science*, *22*, 1220–1226. doi:10.1177/0956797611419302

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

Igelström, K., Webb, T. W., Kelly, Y. T., & Graziano, M. S. A. (2016). Topographical organization of attentional, social and memory processes in the human temporoparietal cortex. *eNeuro*, *3*, doi:10.1523/ENEURO.0060-16.2016 doi:10.1523/ENEURO.0060-16.2016

James, W. (1890). *The principles of psychology*. New York, NY: Henry Holt.

Jiang, Y., Costello, P., Fang, F., Huang, M., & 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. doi:10.1073/pnas.0605678103

Kelly, Y. T., Webb, T. W., Meier, J. D., Arcaro, M. J., & 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. doi:10.1073/pnas.1401201111

Kentridge, R. W., Heywood, C. A., & Weiskrantz, L. (1999). Attention without awareness in blindsight. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *266*, 1805–1811. doi:10.1098/rspb.1999.0850

Kentridge, R. W., Heywood, C. A., & Weiskrantz, L. (2004). Spatial attention speeds discrimination without awareness in blindsight. *Neuropsychologia*, *42*, 831–835. doi:10.1016/j.neuropsychologia.2003.11.001

Kihlstrom, J. F. (2002). Mesmer, the Franklin commission, and hypnosis: A counterfactual essay. *International Journal of Clinical and Experimental Hypnosis*, *50*, 407–419. doi:10.1080/00207140208410114

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

Lambert, A., Naikar, N., McLachlan, K., & 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. doi:10.1037/0096-1523.25.2.321

Lau, H., & Rosenthal, D. (2011). Empirical support for higher-order theories of conscious awareness. *Trends in Cognitive Sciences*, *15*, 365–373. doi:10.1016/j.tics.2011.05.009

Le Meur, O., Le Callet, P., & Barba, D. (2006). A coherent computational approach to model the bottom-up visual attention. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*, 802–817. doi:10.1109/TPAMI.2006.86

Mack, A., & Rock, I. (2000). *Inattentional blindness*. Cambridge, MA: MIT Press.

Mazzoni, P., & Krakauer, J. W. (2006). An implicit plan overrides an explicit strategy during visuomotor adaptation. *Journal of Neuroscience*, *26*, 3642–3645. doi:10.1523/JNEUROSCI.5317-05.2006

McCormick, P. A. (1997). Orienting attention without awareness. *Journal of Experimental Psychology, Human Perception and Performance*, *23*, 168–180. doi:10.1037/0096-1523.23.1.168

Medina, J., & Coslett, H. B. (2016). What can errors tell us about body representations? *Cognitive Neuropsychology*, *33*, 5–25. doi:10.1080/02643294.2016.1188065

Metzinger, T. (2009). *The ego tunnel: The science of the mind and the Myth of the self*. New York: Basic Books.

Molenberghs, P., Johnson, H., Henry, J. D., & Mattingley, J. B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neuroscience and Biobehavioral Reviews*, *65*, 276–291. doi:10.1016/j.neubiorev.2016.03.020

Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, *83*, 435–450. doi:10.2307/2183914

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

Norbre, K., & Kaster, S. (2014). *The oxford handbook of attention*. New York: Oxford Univeristy Press.

Odegaard, B., Knight, R. T., & Lau, H. (2017). Should a few null findings falsify prefrontal theories of conscious perception? *The Journal of Neuroscience*, *37*, 9593–9602. doi:10.1523/JNEUROSCI.3217-16.2017

Parsons, L. M. (1987). Imagined spatial transformations of one's hands and feet. *Cognitive Psychology*, *19*, 178–241. doi:10.1016/0010-0285(87)90011-9

Pattie, F. A. (1994). *Mesmer and animal magnetism: A chapter in the history of medicine*. Hamilton, NY: Edmonston Publishing.

Pesquita, A., Chapman, C. S., & Enns, J. T. (2016). Humans are sensitive to attention control when predicting others' actions. *Proceedings of the National Academy of Sciences, U. S. A.*, *113*, 8669–8674. doi:10.1073/pnas.1601872113

Pettitt, P. (2010). *The paleolithic origins of human burial*. New York: Routledge.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*, 515–526. doi:10.1017/S0140525X00076512

Prinz, W. (2017). Modeling self on others: An import theory of subjectivity and selfhood. *Consciousness and Cognition*, *49*, 347–362. doi:10.1016/j.concog.2017.01.020

Pullman, P. (2000). *The amber spyglass*. New York: Scholastic.

Ramachandran, V. S., & Rogers-Ramachandran, D. (2000). Phantom limbs and neural plasticity. *Archives of Neurology*, *57*, 317–320. doi:10.1001/archneur.57.3.317

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*, 168–185. doi:10.1016/j.neuron.2009.01.002

Rosenthal, D. (1991). *The nature of mind*. New York: Oxford University Press.

Rosenthal, D. (2005). *Consciousness and mind*. New York: Oxford University Press.

Rowling, J. K. (1999). *Harry Potter and the Prisoner of Azkaban*. London: Bloomsbury.

Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: fMRI investigations of theory of mind. *NeuroImage*, *19*, 1835–1842. doi:10.1016/S1053-8119(03)00230-1

Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, *17*, 692–699. doi:10.1111/j.1467-9280.2006.01768.x

Schmidt, K. (2011). Göbekli Tepe: A neolithic site in southwestern Anatolia. In S. R. Steadman & G. McMahon (Eds.), *The Oxford handbook of ancient Anatolia* (pp. 917–933). Oxford: Oxford University Press.

Schwemmer, M. A., Feng, S. F., Holmes, P. J., Gottlieb, J., & Cohen, J. D. (2015). A multi-area stochastic model for a covert visual search task. *PLoS One*, *10*, e0136097. doi:10.1371/journal.pone.0136097

Sekiyama, K. (1982). Kinesthetic aspects of mental representations in the identification of left and right hands. *Perception and Psychophysics*, *32*, 89–95. doi:10.3758/BF03204268

Shadmehr, R., & Mussa-Ivaldi, F. A. (1994). Adaptive representation of dynamics during learning of a motor task. *Journal of Neuroscience*, *14*, 3208–3224. doi:10.1523/JNEUROSCI.14-05-03208.1994

Shulman, G. L., Pope, D. L., Astafiev, S. V., McAvoy, M. P., Snyder, A. Z., & Corbetta, M. (2010). Right hemisphere dominance during spatial selective attention and target detection occurs outside the dorsal frontoparietal network. *Journal of Neuroscience*, *30*, 3640–3651. doi:10.1523/JNEUROSCI.4085-09.2010

Sidky, H. (2017). *The origins of shamanism, spirit beliefs, and religiosity: A cognitive anthropological perspective*. London: Lexington Books.

Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, *28*, 1059–1074. doi:10.1068/p281059

Simons, J. S., Garrison, J. R., & Johnson, M. K. (2017). Brain mechanisms of reality monitoring. *Trends in Cognitive Sciences*, *21*, 462–473. doi:10.1016/j.tics.2017.03.012

Taylor, J. A., Krakauer, J. W., & Ivry, R. B. (2014). Explicit and implicit contributions to learning in a sensorimotor adaptation task. *Journal of Neuroscience*, *34*, 3023–3032. doi:10.1523/JNEUROSCI.3619-13.2014

Thoroughman, K. A., & Shadmehr, R. (2000). Learning of action through adaptive combination of motor primitives. *Nature*, *407*, 742–747. doi:10.1038/35037588

Thoroughman, K. A., & Taylor, J. A. (2005). Rapid reshaping of human motor generalization. *Journal of Neuroscience*, *25*, 8948–8953. doi:10.1523/JNEUROSCI.1771-05.2005

Titchner, E. B. (1898). The feeling of being stared at. *Science*, *8*, 895–897. doi:10.1126/science.8.208.895

Tolkien, J. R. R. (1955). *The lord of the rings*. London: Allen and Unwin.

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

Turing, A. M. (1950). Computing machinery and intelligence. *Mind; A Quarterly Review of Psychology and Philosophy*, *59*, 433–460. doi:10.1093/mind/LIX.236.433

Vallar, G., & 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

Vallar, G., & Ronchi, R. (2009). Somatoparaphrenia: A body delusion. A review of the neuropsychological literature. *Experimental Brain Research*, *192*, 533–551. doi:10.1007/s00221-008-1562-y

van den Boogaard, J., Treur, J., & Turpijn, M. (2017). A neurologically inspired neural network model for Graziano's attention schema theory for consciousness. *International Work Conference on the Interplay Between Natural and Artificial Computation: Natural and Artificial Computation for Biomedicine and Neuroscience*, *10337*, 10–21. Part 1. doi:10.1007/978-3-319-59740-9_2

van Vugt, B., Dagnino, B., Vartak, D., Safaai, H., Panzeri, S., & Dehaene, S. (2018). The threshold for conscious report: Signal loss and response bias in visual and frontal cortex. *Science*, *360*, 537–542. doi:10.1126/science.aar7186

Webb, T. W., & Graziano, M. S. A. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, *6*, doi:10.3389/fpsyg.2015.00500

Webb, T. W., Igelström, K., Schurger, A., & Graziano, M. S. A. (2016a). Cortical networks involved in visual awareness independently of visual attention. *Proceedings of the National Academy of Sciences, U. S. A.*, *113*, 13923–13928. doi:10.1073/pnas.1611505113

Webb, T. W., Kean, H. H., & Graziano, M. S. A. (2016b). Effects of awareness on the control of attention. *Journal of Cognitive Neuroscience*, *28*, 842–851. doi:10.1162/jocn_a_00931

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, *15*, 728–755. doi:10.1080/17405629.2018.1435413

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*, 103–128. doi:10.1016/0010-0277(83)90004-5

Winer, G. A., Cottrell, J. E., Gregg, V., Fournier, J. S., & Bica, L. S. (2002). Fundamentally misunderstanding visual perception: Adults' belief in visual emissions. *American Psychologist*, *57*, 417–424. doi:10.1037/0003-066X.57.6-7.417

Yeo, B. T. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., … Buckner, R. L. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of Neurophysiology*, *106*, 1125–1165. doi:10.1152/jn.00338.2011