Michael S.A. Graziano

Attributing Awareness to Others

The Attention Schema Theory and Its Relationship to Behavioural Prediction

Abstract: *The attention schema theory provides a single coherent framework for understanding three seemingly unrelated phenomena. The first is our ability to control our own attention through predictive modelling. The second is a fundamental part of social cognition, or theory of mind — our ability to reconstruct the attention of others, and to use that model of attention to help make behavioural predictions about others. The third is our claim to have a subjective consciousness — not merely information inside us, but something else in addition that is non-physical — and to believe that others have the same property. In the attention schema theory, all three phenomena stem from the same source. The brain constructs a useful internal model of attention. This article summarizes the theory and discusses one aspect of it in greater detail: how an attention schema may be useful for predicting the behaviour of others. The article outlines a hypothetical, artificial system that can make time-varying behavioural predictions about other people, and concludes that attributing some form of awareness to others is a useful computational part of the prediction engine.*

Correspondence:
Michael S.A. Graziano, Department of Psychology and Neuroscience,
Princeton University, Princeton, NJ 08544, USA.
Email: graziano@princeton.edu

Journal of Consciousness Studies, **26**, No. 3–4, 2019, pp. 17–37

1. Introduction

The attention schema theory of consciousness (AST) was first described eight years ago, and has been elaborated in many publications since (e.g. Graziano, 2010; 2013; 2016; Graziano and Kastner, 2011; Webb and Graziano, 2015). In the present article, I will briefly summarize the theory and then provide a more focused discussion of one part that has not been explored in as much depth.

In previous accounts, the theory was discussed in terms of its potential for explaining how the human brain arrives at the claim of subjective consciousness in the first place, the relationship between consciousness and attention (Webb and Graziano, 2015), and the relationship between consciousness and social cognition (Graziano, 2013; Kelly *et al.*, 2014). Here I will focus on how the theory intersects one specific part of social cognition, predicting the behaviour of other people. Behavioural prediction has been discussed in many contexts, and depends heavily on theory of mind, or the ability to attribute beliefs, intentions, emotions, goals, and agendas to others (Baron-Cohen, 1997; Frith and Frith, 2003; Premack and Woodruff, 1978; Saxe and Kanwisher, 2003; Wimmer and Perner, 1983). One component of theory of mind, modelling the attention of others, however, is often treated in a superficial manner, for example conflating visual attention with the simple direction of gaze. Through a thought experiment, I will discuss how a good behavioural prediction machine should build a model of other people's attention, and how constructing a model of someone's attention may be tantamount to attributing subjective awareness to the person. We never do attribute true, mechanistic attention to other people — we never look at a person and say to ourselves, 'Ah, his neurons are using lateral inhibitory processes to provide a competition among signals, which is biased by internal directives, causing the signals related to that doughnut to dominate the brain's computational processes and enhance the likelihood of a behavioural reaction…' Instead, we build a simpler, schematic model. We attribute to the other person a more vague property of awareness that has currently seized on the doughnut, but has maybe failed to seize on the bug crawling up his sleeve. Socially, we seem to use awareness as a model of attention. Without that model of the attention that others direct to the objects around them, it is extremely difficult or impossible to predict their behaviour. With an attention schema, or a simplified model of attention, behavioural prediction is enabled.

I will also argue below that an attention schema may help to explain some of the most common, physically irrational intuitions that people have about an invisible, energy-like essence inside us that can sometimes emanate from the eyes. The power of the AST lies in its ability to explain why people have simplified, schematized models of our own internal processes, giving us physically incoherent intuitions about what goes on inside our own heads. Evolution is under no obligation to supply us with built-in, internal models that are scientifically accurate in all their details. Instead, it builds internal models that are useful, and yet often cut corners for efficiency.

2. A Brief Summary of the Theory

This article does not provide a complete account of the AST. It does not itemize and answer the many common questions and concerns, it does not describe the relationship between the theory and specific networks in the cerebral cortex, and it does not present the growing set of experimental studies supporting specific predictions of the theory. The reader is referred to other sources for a more complete and a more data-oriented treatment (e.g. Graziano, 2013; Kelly *et al.*, 2014; Webb and Graziano, 2015; Webb, Kean and Graziano, 2016; Webb *et al.*, 2016). Instead, in this section, I summarize the general idea of an attention schema in order to motivate a thought experiment about behavioural prediction described in the following sections.

The AST can be summarized in three broad points. First, the brain is an information processing machine. Second, it has a capacity to focus its processing resources more on some signals than on others. That focus may be on select, incoming sensory signals, or it may be on internal information such as specific, recalled memories or emotional states. That ability to process select information in a focused and deep manner is sometimes called attention. Third, the brain not only uses the process of attention, but it also builds a set of information — a representation, or an internal model — descriptive of attention. That internal model is the attention schema.

In the theory, the attention schema provides the requisite information that allows the machine to make claims about consciousness — to claim that it has a subjective awareness of something. Logically, any claim that the machine makes must be based on information contained within it. This theory proposes that the source information for the claim of subjective awareness is an attention schema.

For example, suppose a person looks at an apple. When the person reports, ‘I have a subjective experience of the apple’, three items are included in that claim: the self, the apple, and a subjective experience that links the two. The claim about the presence of a self depends on cognitive access to a deeper, self-model. Without a self-model, without the requisite information, the system would be unable to make claims referencing the self. The claim about the presence of an apple depends on cognitive access to a model of the apple, presumably constructed in the visual system as one looks at the apple. Again, without the requisite information, the system would obviously be unable to make any claims about the apple or its visual properties. Finally, in the theory, the claim about the presence of subjective experience depends on cognitive access to an internal model of attention. That internal model describes the process of attention itself — about how the brain is focusing resources. It does not provide a scientifically precise description of attention, complete with the details of neurons, lateral inhibitory synapses, and competitive signals. The model is silent on the physical mechanisms of attention. Instead, it provides a simplified and schematic description of some of the main dynamics and consequences of attention. The heart of the AST is the proposal that if you could provide a description of attention — not of the thing you are attending to, but of the act of attention itself — while leaving out the mechanistic details of neural implementation that the brain has no need to know about, the description would match the property of conscious experience that we claim to have.

In a typical intuitive account of consciousness, one does not just process the information that the apple is red — one experiences redness. Redness has a ‘what it feels like’ aspect. The experience itself has few overt physical attributes. Experience cannot be measured in grams on a scale, it does not occlude light, it has no definite height or width. Yet it is presumed to exist. It is, literally, a non-physical thing. It exists outside of the normal dimensions of physicality, and in this sense is metaphysical. In this article, when I refer to a non-physical or metaphysical essence of consciousness, it is to this experiential component I am referring. The heart of the problem of consciousness research has been: how can physical states in the brain cause this non-physical essence, this subjective, experiential adjunct to brain activity?

One of the central contentions of the AST is that everything we know about the world and ourselves, everything we believe and everything we claim, no matter how intuitively obvious it seems or how fervently we claim it, derives from information in the brain. Our deep, internal models, packets of information descriptive of ourselves and our world, provide us with what we think is our reality. Higher cognition accesses those deep, internal models, and reports their content as though it were literally true. Given the three internal models discussed above concerning the self looking at an apple, cognitive machinery can produce three general types of claims. First, it can make claims about the self (based on information in the self-model). Second, it can make claims about the apple (based on information in the internal model of the apple). Third, it can make claims about a mental experience or possession that the self has of the apple. This last claim is based on information in the attention schema — information that is superficially descriptive of attention.

For example, the machine might claim that the mental possession of the apple — the mental possession in and of itself — has few describable physical properties (since the attention schema lacks information on the physical mechanism of attention). According to the information available to the machine, that mental possession has no weight, opacity, colour, hardness, smell, sound, or working parts. It does, however, have at least one physical attribute: a general location somewhere inside the body. It may have other vaguely physical properties, such as an energy-like capacity to cause things to happen — to make us choose and act. In that superficial, but useful, schematic description of attention, a weird internal essence enables one to understand the apple, to grasp the details of the apple in vivid form, to react to and to remember the apple. In the AST, the brain, in relying on the partial and schematic information contained in its attention schema, claims to have a consciousness of the apple.

The AST is not a traditional theory of consciousness. It does not explain how a physical brain produces that illusive, non-physical experience. Instead, it explains how a machine *claims* to have a non-physical experiential essence, and how it cannot tell that the claim is based on computations and internal models.

The AST is similar to the 'higher-order representational' approach to consciousness. In that approach, we are conscious of a mental state when the brain builds a second-order representation of that state. The most prominent current example of the representational approach is the higher-order thought theory (e.g. Gennaro, 1996; Carruthers, 2000; Rosenthal, 2006). To be conscious of the apple is not merely to build a representation of the apple, but to build a representation of the fact that you are processing the apple. The AST could be considered a specific example of the higher-order thought theory. However, arguably the higher-order thought theory, at least as it is often described, does not explain how the machine claims to have an experience, a 'what it feels like' component, the subjective feel of the redness, roundness, and shininess that make up the apple. It explains how a machine 'knows' about the apple, and also 'knows' that it is processing the apple. But why exactly would a system like that claim to have a subjective phenomenon attached to the apple? The AST specifies the origin of that claim. In the theory, a representation of attention provides the crucial information that leads to the claim that 'There is such a thing as a subjective experience, a what-it-feels-like, and right now that feeling is attached to the red shiny apple'.

The AST is also consistent with the perspective called illusionism (e.g. Dennett, 1991; Metzinger, 2009; Humphrey, 2011; Hood, 2012; Frankish, 2016; Kammerer, 2016; Blackmore, 2016). The AST is particularly close to Dennett's work (Dennett, 1991). However, I tend not to use the term illusionism when describing the theory. My concern is not with the underlying concepts of illusionism, which seem sound to me, but with the colloquial connotations that often come along with the word itself. I find the word awkward for three main reasons.

First, in my experience, when I speak to people from varied backgrounds, they tend to equate an illusion with a glitch in the system that at best should be ignored, and at worst is harmful. If we can see through the illusion, we are better off. Yet in the AST, the attention schema is a well-functioning internal model. It is not normally dysregulated or in error.

Second, when the word is used colloquially or metaphorically, it usually indicates that something appears to be present but actually does not exist. For example, if I were complaining to you and said, 'I swear, my boss's competence is an illusion', I don't mean, 'He's competent, but in a slightly different way from what you might expect, like a straw that looks slightly bent in water'. When used colloquially rather than technically, that phrase means, 'His competence doesn't exist'. If consciousness is an illusion, then by the colloquial implication of that phrase, nothing real is present behind the illusion. There is no 'there' there. But in the AST, that is not so. Consciousness is a good, if detail-poor, account of something real: attention. We do have attention, a physical and mechanistic data-handling process that emerges from the interactions of neurons. When we claim to have consciousness, we are providing a slightly schematized version of the literal truth. There is, indeed, a 'there' there. In the

AST, one might say that consciousness is more like a caricature than an illusion.

Third, calling consciousness an illusion, in my view, boxes one in the wrong philosophical arena. The AST is not a theory of how the brain constructs experiences, illusory or otherwise. One does not want to get caught having to explain: if consciousness is an illusion, what is experiencing the illusion? Or, if nothing is experiencing the illusion, why is it called an illusion? The AST is a theory of how a machine constructs information and makes claims — how it claims to have experiences — and being stuck in a logic loop, or captive to its own internal information, it cannot escape making those claims. At its heart, the AST is not a philosophical theory. It is an engineering theory of how a machine works.

The AST's explanation of consciousness is, in some ways, the least important part of the theory. In the AST, the mechanism that lies behind consciousness, the attention schema itself, plays several fundamental, adaptive roles in brain function. These roles go far beyond merely allowing us humans to walk around bragging about a metaphysical inner life.

One possible adaptive function of an attention schema is to help control attention (Webb and Graziano, 2015). A fundamental principle of control theory is that a good controller should incorporate an internal model (Conant and Ashby, 1970; Francis and Wonham, 1976; Camacho and Bordons Alba, 2004). Much like a self-driving car needs an internal model of the car, or the motor system in the brain relies on an internal model of the arm, so the brain's controller of attention should incorporate an internal model of attention — a set of information that is continuously updated and that reflects the dynamics and the changing state of attention (Webb and Graziano, 2015; Webb, Kean and Graziano, 2016). Since attention is one of the most important processes in the brain, the proposed attention schema, helping to control attention, would be of fundamental importance to the system.

A second proposed adaptive function of an attention schema is to contribute to social cognition — using the attention schema to model the attentional states of others (Graziano, 2013; Kelly *et al.*, 2014). A main advantage of this social use of an attention schema lies in behavioural prediction. How can I predict your behaviour? Whatever item you are attending to, you are likely to behave toward, and what you are not attending to, you are much less likely to behave toward. If

I have a basic model of attention, of its dynamics and consequences, then I can make better predictions about your behaviour.

In the next section, I ask: if we try to build an artificial, behavioural prediction engine, will we find it useful to include an attention schema, and what properties might that attention schema have? At least some initial work has been done on building machines that can perform social cognition (e.g. Baker, Saxe and Tenenbaum, 2009; Rabinowitz *et al.*, 2017; Saxe and Houlihan, 2017; Yoshida, Dolan and Friston, 2008). Most of these previous attempts focus on reconstructing other people's motivations, intentions, or beliefs — components of a traditional theory of mind. In the following section, I focus on a simpler but still fundamental component to theory of mind, modelling the attentional states of others.

3. A Behavioural Prediction Engine: A Thought Experiment

A man walks into a small room. Unseen, a camera eye watches him and a microphone records sounds in the room. These devices are connected to an artificial system whose job is to predict his moment-by-moment unfolding behaviour. How can we build that prediction engine?

The room contains the following three items. First: a white powdered doughnut in the middle of a table, in the middle of the room. The overhead light shines brightest on the doughnut. Second: a small puddle of water on the floor in front of the table. If he walks to the table and is not careful, he'll step in the puddle. Third: a phone on a shelf in the corner of the room, where the light is dim.

The first task of the prediction engine is to identify the affordances in this environment into which the person has just walked. Gibson (1979) coined the term affordance to refer to an aspect of an agent's environment that provides an opportunity for action. Agents perceive their environments partly in terms of these action opportunities. The actions are often ethologically meaningful, or in some way specific to the animal species. A fly provides an affordance to a frog — grabbing with the tongue and eating. A branch provides an affordance to a bird — perching. A doorknob provides an affordance to a person — grasping and turning. In the Gibsonian sense, an affordance refers to how an agent perceives its own opportunities for actions. In the case of our prediction engine, however, we require it to function in a third-person manner, identifying the affordances relevant to the man as he enters the room.

We can already see the great complexity of building a working, behavioural prediction engine. Each object can occasion an extremely large number of affordances. For example, with respect to the puddle, the person might step over it; he might choose to jump into it and make a splash; he might take a paper towel out of his pocket and mop it up. All of these are possible intentional behaviours toward the object. As for the doughnut, the man might reach for it and eat it; he might hold it to his eye and pretend it's a monocle; he might throw it on the floor and stomp on it. As for the phone, he might pick it up and try to unlock it, or he might pocket it surreptitiously. The prediction engine is faced with a large set of possible affordances, even in a room with a minimalist collection of objects.

For the purposes of the present discussion, we will give our machine a head start. Let us assume this difficult problem has already been solved. We will simply hand our prediction engine a complete list of all affordances relevant to the three items in the room.

We will do even more for our prediction machine. We will add in the whole apparatus of a traditional theory of mind (Baron-Cohen, 1997; Frith and Frith, 2003; Premack and Woodruff, 1978; Saxe and Kanwisher, 2003; Wimmer and Perner, 1983). Let's simply assume our machine already can attribute to the man beliefs, desires, emotions, and goals. The machine is already equipped with the probability that, having walked into that room, the man will do each of the identified actions. In general, most people do not jump into a puddle or stomp on a doughnut. These are low probability events. Eating the doughnut is higher probability. Moreover, some probabilities are specific to the particular person at the time he enters the room. For example, if I know he hasn't eaten for ten hours, I might suppose a higher probability of him eating that doughnut. If I know he has diabetes, I might suppose he won't eat it. If I know he has an impulse control problem, or is angry, I might boost the probability of him stomping that doughnut. We will not task our machine with the complexities of computing these many background states, often called a theory of mind, or, as it has also been called, the intentional stance (Dennett, 1987). Let us suppose we can take available statistical information on general human behaviour, combine it with estimated information on this specific person, and roll it into a list of numbers. In that list, each affordance has a probability attached to it, which we will call the prior probability, P_{prior}. The prior probability is an estimate of how the man will act on entering that room. Maybe there is a 30% probability he'll eat the doughnut, a 20% probability he'll step over the puddle, and so on, down the list of affordances. We seem to have done all the work and given the crucial information to the prediction machine. Is there anything left for that lazy machine to do?

Even with all of that useful information front-loaded into it, the machine still cannot predict the man's behaviour on a moment-by-moment basis. The machine needs information about a crucial, hidden variable: how the man is focusing his processing resources on his environment. His processing resources are constantly shifting, moving about his environment in complex, changing patterns. What he is attending to, he is more likely to react to at that moment, and what is outside his attention at any moment, he is extremely unlikely to react to. As a result, the probabilities for the many affordances are constantly in flux.

Take the case of the doughnut on the table. Suppose the artificial system, with its front-loaded, theory-of-mind information, already knows that the man likes doughnuts, and is hungry, and has a certain estimated probability of picking up the doughnut and eating it. Recall that we called this initial, estimated probability for that particular act, P_{prior}. Now suppose that the prediction machine computes a time-varying factor, $C_1(t)$. C_1 varies between 0 and 1, and represents a normalized measure of the amount by which the man is focusing his processing resources on object 1, the doughnut. The more his attention is focused on the doughnut, the more likely he is to act toward it. It is a kind of permissive variable, permitting the possibility of action. Now I will present the only equation in this article, with apologies both to those who prefer more equations and those who don't like any. Suppose the machine makes a simple calculation: the probability that the man will eat the doughnut is:

$$P_{action} = C_1 \text{ x } P_{prior}.$$

As C_1 varies in time, the computed probability of the man engaging in that specific action changes. Most of the time, the man is paying little or no attention to the doughnut. C_1 is close to 0, therefore $P_{action} = 0$, and the machine predicts he won't eat it. Occasionally his attention to the doughnut may flicker up, and C_1 will rise. His attention to the doughnut may even surge to a maximum, and then C_1 will temporarily peak around 1. As soon as his attention to the doughnut spikes, his estimated probability of eating it spikes. Even at the peak of that spike, his probability of eating the doughnut never exceeds P_{prior}, which may after all be quite small, since people don't often pick up and eat random doughnuts. As his attention to the doughnut drops back down again, his probability of eating it also subsides back toward zero. In this manner, his probability of reaching out for the doughnut to eat it fluctuates moment-by-moment in a way the machine can track. The usefulness of this kind of computation is to take the more standard theory-of-mind approach, which tends to operate in the framework of static vignettes, and put it in a framework where dynamic and sometimes drastic changes of attention from second to second can be accommodated.

The task of our prediction engine is to compute C_1, the amount by which the man is focusing processing resources on object 1. But it is a difficult task. The prediction machine does not have direct access to the man's brain. Even if it somehow did — if the machine could insert millions of electrodes and monitor the internal neural processes — it would then face the impossibly complicated task of reconstructing and modelling the actual physical, neural interactions that compose attention. Moreover, there are many different overlapping kinds and layers of attention — exogenous, endogenous, spatial, feature, attentional switching, inhibition of return, just to give a few examples. How can our machine reconstruct the tangled, massively multi-component truth of that man's attentional processes? Instead, the prediction engine needs a much simpler, schematized model of attention that can be constrained by sparse observation.

Here I will outline three examples of the heuristics that a machine might use to estimate the man's attention. To clarify, I am not claiming to present new insights about how attention works. I am mining established scientific insights, to help cobble together a reasonable working model of attention for our hypothetical behavioural prediction machine to use.

One heuristic is gaze. Treating gaze as a proxy for visual attention has a long history in psychology and neuroscience (e.g. Calder *et al.*, 2002; Friesen and Kingstone, 1998; Hoffman and Haxby, 2000; Kobayashi and Koshima, 1997; Perrett *et al.*, 1985; Baron-Cohen, 1997). If the man's gaze is directed at or near the doughnut, his processing resources are more likely to be focused on the doughnut. However, it is important to realize that gaze is only one, imperfect cue. Gaze and attention are not the same. The man could be staring straight at the doughnut and yet occupied by something else, covertly concentrating on a nearby object, listening intently to a nearby sound, occupied by an itch on his arm, or thinking hard about his plans for tomorrow. But, on average, gaze is still a useful, if probabilistic, consideration when trying to estimate the man's attention. If his gaze is on X, then the machine might estimate a high value of C_1.

A second heuristic is salience. The doughnut is in a central location, it is white, and it is under a bright light. It has high stimulus salience against the background, which tends to increase attention. The puddle on the floor and the cell phone on a back shelf, in contrast, have low stimulus salience. Given this heuristic, our machine should set C_1 to a high value.

A third heuristic is competition. The doughnut is alone on an otherwise empty table, and attention depends inversely on clutter or visual competition. This principle of competition will become especially relevant below as I discuss the man's possible behaviour toward the other two objects in the environment, the puddle on the floor and the cell phone on the shelf.

Many other heuristics may be useful as well. These heuristics are taken straight from the basic, current knowledge in cognitive psychology about the dynamics and properties of attention. Given these heuristics and sparse clues, the machine can observe the room, observe the man, and estimate a time-varying value for C_1, the amount of processing resources the man is directing at object 1. The machine can then estimate the fluctuating probability that the man will engage in a specific action with respect to the doughnut.

Now consider the second object, the puddle on the floor. At each moment in time, will the man react to the puddle, for example stepping over it? The prediction machine must compute a time-varying value for C_2, which represents the amount by which the man is focusing his processing resources on object 2. Note that the value of C_2 interacts with the value of C_1. Because the man's processing resources are limited, as C_1 increases, C_2 must decrease. Our prediction engine must take into account the competitive dynamics of attention. In other words, if he seems to be highly attentive to the doughnut as he walks toward the table, there is a high chance, at that moment, that he'll walk right into the puddle without stepping over it. This is a simple but effective behavioural prediction, that seems intuitively obvious to any normal person, but that depends on an internal model of the dynamics of attention.

The machine can also compute C_3, an estimate of how much the man's processing resources are focused on the third object, the phone. Initially, the machine computes a low value for C_3 because the phone is not a salient object and the man's gaze is not directed at it. Now the phone rings. That object suddenly gains higher salience. Registering that change in salience, the machine can compute a sharp increase in C_3. Even if the man's gaze is fixed elsewhere, the machine can still compute that, given the intense salience of the stimulus, C_3 is likely to be high. As a direct consequence, C_2 and C_1 must drop at that moment in time. In the moment the phone rings, the probability that he'll reach for the doughnut dips. In colloquial parlance, he's been momentarily distracted. People understand this behavioural prediction intuitively, but again, that intuition depends on a model of the dynamics of attention. Moreover, the man's attention, drawn to the phone when it rings, has some stickiness or viscosity. It will tend to remain focused on the phone for some period of time that is typical or characteristic of human attention, perhaps half a second, before it can move away. Thus the computation of a time-varying C_3 must depend partly on an approximate model of the sluggishness or viscosity of visual attention.

To summarize this example of the man in the room, the prediction engine watching him is constantly computing an ever-changing vector C, whose components are C_1, C_2, and C_3. Based on those values, it can estimate the probability that the man will engage in actions that are afforded by the doughnut, the puddle, and the phone. That computation is based on a rich model of how attention works — how the man's processing resources are deployed in real time. To be useful, the model must incorporate factors such as where the man's eyes are directed, the salience of stimuli in the environment, the clutter or competition in the environment, competition between the three likely objects of attention, and the temporal dynamics, such as the viscosity, of normal human attention.

In effect, the prediction engine has constructed a simple, cartoonish model of the man's attention. In that model, the man has a property — call it Substance C. That substance is invisible — it cannot be directly observed. It does not, itself, register on the machine's camera. It has no physical texture, no opacity, no hardness, no sound. Substance C has its source inside the man and flows out, with a bias toward flowing out of the eyes along straight lines, although it is not always in lockstep with the eyes. It makes contact with specific objects in the environment. It is slightly viscous, in the sense that it adjusts sluggishly as the flow is redirected from one object to another. It is a limited resource, in the sense that if more is flowing toward one object, less is flowing toward other objects. It can be partitioned among objects, but tends to be directed mainly at one object at a time. It also has an energy-like or will-like property, in the sense that when it flows from the agent to the object, it empowers the agent. It does not directly galvanize the agent to act; nor does its presence determine the specific action; instead its presence empowers the agent to make a behavioural choice about whether to act and what act to employ.

Note that Substance C behaves like a classical fluid in the following formal ways. It springs from a source. It can flow out from the source and changes direction with some viscosity. The total amount of fluid is conserved — meaning that, like spraying water out of a hose, if you spray a lot at one object, you can't at the same time spray a lot at a second object. Substance C, that invisible viscous fluid, is a construct of the prediction engine. It is a useful proxy for the neuronal processing occurring within the agent. The reality is that the man's brain contains interacting networks of neurons, whereas the prediction engine attributes to the man something very similar to a metaphysical consciousness stuff that streams out of the eyes.

Substance C is a simplified version of attention. It is an attention schema, tailored for modelling someone else's attention.

I suggest that we humans are prediction engines of this sort. We are constantly attributing to each other a Substance C. Attributing to agents an invisible, metaphysical essence of consciousness is a useful — I would argue fundamental — component of behavioural prediction. It is not the only component. Note that this hypothetical prediction machine would fail utterly if it contained only an attention schema. The model of attention works because the machine also contains other aspects of theory of mind. The machine needs rich information on the man, attributing to him beliefs, emotions, and agendas. By adding a model of attention, we enable the machine to make behavioural predictions that are sensitive to moment-by-moment changes in attention, as the man processes the world around him.

The simplified discussion above involves concrete objects such as doughnuts and puddles. However, we are also capable of attributing to each other an awareness or consciousness of intangible, abstract thoughts, memories, or emotions. If, as I suggest, we model attention as something like an invisible substance that flows from a person to an object, then how do we model attention to an internal event? For example, people can attend to the thought that 2 + 2 = 4. When they do so, they withdraw attention from external, sensory events. In that case, the 'Substance C' is not flowing out of the eyes to an external object, but instead is entirely internal to the person. It is contained within the head, moving among ideas rather than interacting with objects in the external world. Because of this mixture of external and internal targets of attention, the real-life case is more rich than the example above, but the principles are the same.

It is worth noting here that the prediction engine may not do a very good job. The man in the room may act in non-predicted ways. He may simply walk around the room doing nothing in particular, muttering to himself, a behaviour stream that is not very amenable to moment-by-moment prediction. His reactions to the objects in the room may be chaotic. I do not think we humans are, on any absolute scale, good at predicting each other's moment-by-moment behaviour. But if the predictions are at all better than chance, they confer a useful advantage.

4. An Implicit Belief in Beams Emanating from the Eyes: Behavioural Evidence

We recently conducted an experiment in my lab on how people implicitly perceive the gaze of others (Guterstam *et al.*, 2018). Participants looked at a computer image of a paper tube standing upright on a table. The participants were asked to imagine the tube being gradually tilted, and to judge the critical angle at which it would probably fall over. With arrow keys on a keyboard, the participants marked out the estimated, critical tilt angle over multiple trials. At the same time, on every trial, a face appeared in the picture. Participants were given no explanation for the face: it was simply present, either on the far left or right side, in profile view, looking directly at the paper tube. In a post-test survey, none of the subjects correctly guessed why the face was present, or thought that the face altered their tilt judgments in any specific manner. And yet the face did have a significant affect on their tilt judgments. It was as if participants perceived beams of energy coming out of the face's eyes, pushing on the paper tube, influencing its critical tilt angle. When the tube was tilting toward the face, the eyes seemed to prop it up, and people judged that it could be tilted further before falling over. When the tube was tilting away from the face, the eyes seemed to give it an extra nudge, and people judged that it would fall over sooner, at a shallower angle. The effect was small, about half a degree of tilt angle, as if the effect of the eyes was similar to a gentle breeze. We also tested several control conditions. In one control, the face in the picture was blindfolded. In another, the face had open, visible eyes, but was facing away from the tube. In a third control, participants were told that although the face was aimed at the tube, it was looking past the tube at the farther wall. In all of these control conditions, the effect went away. The estimated critical tilt angle was the same whether the tube was angled toward or away from the face, as though eye beams were no longer affecting the tube.

This apparent effect of eye beams was not explicit. In a post-test survey, we asked the participants how they thought vision worked: did it involve something coming out of the eyes, or something going in? Only five percent of subjects, evidently with a poor science education, explicitly reported a belief in beams coming out of the eyes. The rest correctly indicated that vision works by light entering the eyes. And yet, at an implicit level, they all seemed to be treating an open eye as though an invisible substance naturally flowed out of it and interacted with the physical world. When the data were restricted to the participants who understood the correct optics, the same implicit effect on tilt judgment was found.

In my interpretation, we were tapping into Substance C. We were observing a simplified, implicit model of visual attention at work. In that view, not only do we attribute the property of consciousness to others — a cartoonish depiction of attention, in which a mind can take subjective, experiential possession of an item — but that model of attention also has a spatial, geometric component to it. We implicitly model consciousness as something that can flow through space from a conscious source. The experiment, to me, highlights the manner in which the brain constructs useful, but simplified, and sometimes physically wrong, models to help it monitor and predict its world.

The extramission theory of vision, in which something invisible emanates from the eyes and physically affects objects in the world, dates back at least to the ancient Greek philosophers (Gross, 1999). The correct theory was not fully worked out until the ninth century AD, when the Arab scientist Ibn al-Haytham studied the laws of optics and realized that light enters the eye in straight lines and forms an image.

A folk belief in eye beams continues to be culturally common. For example, a belief in an ‘evil eye’ is still widespread (Dundes, 1981). In our own culture, Superman has beams that can emanate from his eyes and burn holes. The terminator robot has red lights in its eyes. We refer to the light of love and the light of consciousness in someone’s eyes, and we refer to death as the moment when light leaves the eyes.

The belief that someone else’s gaze can touch or press on another person was so widespread that, more than a hundred years ago, Titchner thought it was worth testing in the lab (Titchner, 1898). In controlled experiments he found, not surprisingly, that people cannot directly feel each other's stares. Despite the lack of a physical basis for it, the belief that vision involves something beaming out of the eyes is so intuitive that it is the default belief among children (Piaget, 1979). A series of studies from the 1990s suggested that many US college students believe the incorrect, extramission theory of vision (Cottrell and Winer, 1994; Winer, Cottrell and Karefilaki, 1996; Winer *et al.*, 2002). In our own study, we found about a 5% rate of belief in extramission which extended across all age groups tested from 18 to 60, and across educational levels from high school to masters degrees (Guterstam *et al.*, 2018).

The reason why these clearly wrong beliefs have so much cultural traction may be that they tap into a deep, automatic, implicit model that may have evolved over millions of years. The model helps us to keep track of other people's attention in an efficient, schematic way, so that we can better predict their behaviour. Even when we know better intellectually, we cannot help that intuition. We not only attribute an awareness stuff to others, we cannot help implicitly taking into account beams of that awareness stuff coming out of them.

5. Modelling Others versus Modelling Self

It is worth making one final point. This article focuses on attributing consciousness to others. I argue that we not only attribute a subjective experience to others, but we also implicitly treat consciousness as a substance that flows outward from a source inside an agent, and that we do so because that model is fundamentally useful in predicting the behaviour of others. What about one's own consciousness? Is it a matter of applying the same process of behavioural prediction to oneself, or does one's own consciousness contain additional layers? How do we arrive at the conviction: 'Not only is that apple red, but I have a subjective, conscious *experience* of the redness!' And do we also implicitly treat our own consciousness as an ethereal substance that can flow outward and touch items in the world?

I have argued in other places (Graziano, 2013) that the social attribution and self-attribution of consciousness are similar, but not identical. 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 regulate one's own behaviour. It may be part of the machinery for the control of one's own attention. 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. All of these considerations suggest to me there are likely to be substantial differences between social attribution and self-attribution of consciousness, built on top of a core similarity. The consciousness we attribute to others may be more like a pale version of the consciousness we attribute to ourselves.

A common misconception about the AST is that in it, attributing awareness to someone else is primary in evolution and development, and only secondarily do we turn that skill inward and construct our own awareness; in effect, consciousness emerges first from social cognition. But that interpretation is a misreading of the theory. In its simplest form, the theory states only that we attribute awareness to agents because that attribution makes for a useful, simplified model of attention. The theory is agnostic about which came first, attributing awareness to oneself or to others. As I have written before (Graziano, 2014), my own guess is that the brain probably evolved a self-model first. The roots of the attention schema seem more likely to lie in modelling, predicting, and controlling one's own attention, a process that must have been relevant in some form at least as far back as half a billion years ago with the emergence of the vertebrate brain. In that interpretation, the brain then secondarily evolved the ability to use an attention schema socially, to model the attentional states of others. However, one could very plausibly construct the opposite hypothesis, that socially attributing consciousness to others came first, as in Prinz's 'import theory' (Prinz, this issue).

Whichever came first, the social attribution of awareness or one's own awareness, in the AST the two are related at a deep level. The power of the AST lies in its ability to link together three classes of phenomena into a single explanatory framework. The first is our ability to internally monitor and control our own attention, through being able to model it predictively. The second is social cognition — especially our ability to model the attention of others, and to use that model to make behavioural predictions. The third is our flamboyant human trait of claiming to have a semi-magical inner state — consciousness.

References

Baker, C.L., Saxe, R. & Tenenbaum, J.B. (2009) Action understanding as inverse planning, *Cognition*, **113**, pp. 329–349.

Baron-Cohen, S. (1997) *Mindblindness: An Essay on Autism and Theory of Mind*, Cambridge, MA: MIT Press.

Blackmore, S. (2016) Delusions of consciousness, *Journal of Consciousness Studies*, **23** (11–12), pp. 52–64. Reprinted in Frankish, K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.

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**, pp. 1129–1138.

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

Carruthers, P. (2000) *Phenomenal Consciousness: A Naturalistic Theory*, Cambridge: Cambridge University Press.

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**, pp. 89–97.

Cottrell, J.E., Winer, G.A. (1994) Development in the understanding of perception: The decline of extramission perception beliefs, *Developmental Psychology*, **30**, pp. 218–228.

Dennett, D.C. (1987) *The Intentional Stance*, Cambridge, MA: Bradford Books/MIT Press.

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

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**, pp. 457–465.

Frankish, K. (2016) Illusionism as a theory of consciousness, *Journal of Consciousness Studies*, **23** (11–12), pp. 11–39. Reprinted in Frankish, K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.

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

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

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

Gibson, J.J. (1979) *The Ecological Approach to Visual Perception*, Boston, MA: Houghton Mifflin Harcourt.

Graziano, M.S.A. (2010) *God, Soul, Mind, Brain: A Neuroscientist's Reflections on the Spirit World*, Teaticket, MA: Leapfrog Press.

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

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

Graziano, M.S.A. (2016) Consciousness engineered, *Journal of Consciousness Studies*, **23** (11–12), pp. 98–115. Reprinted in Frankish, K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.

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

Gross, C.G. (1999) The fire that comes from the eye, *The Neuroscientist*, **5**, pp. 58–64.

Guterstam, A., Kean, H.H., Webb, T.W., Kean, F.S. & Graziano, M.S.A. (2018) An 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, USA*, **116** (1), pp. 328–333.

Hoffman, E.A. & Haxby, J.V. (2000) Distinct representations of eye gaze and identity in the distributed human neural system for face perception, *Nature Neuroscience*, **3**, pp. 80–84.

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

Humphrey, N. (2011) *Soul Dust*, Princeton, NJ: Princeton University Press.

Kammerer, F. (2016) The hardest aspect of the illusion problem — and how to solve it, *Journal of Consciousness Studies*, **23** (11–12), pp. 124–139. Reprinted in Frankish, K. (ed.) (2017) *Illusionism as a Theory of Consciousness*, Exeter: Imprint Academic.

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 USA*, **111**, pp. 5012–5017.

Kobayashi, H. & Koshima, S. (1997) Unique morphology of the human eye, *Nature*, **387**, pp. 767–768.

Metzinger, T. (2009) *The Ego Tunnel: The Science of the Mind and the Myth of the Self*, New York: Basic Books.

Perrett, D.I., Smith, P.A., Potter, D.D., Mistlin, A.J., Head, A.S., Milner, A.D. & Jeeves, M.A. (1985) Visual cells in the temporal cortex sensitive to face view and gaze direction, *Proceedings of the Royal Society of London B: Biological Sciences*, **223**, pp. 293–317.

Piaget, J. (1979) *The Child's Conception of the World*, Tomlinson, J. & Tomlinson, A. (trans.), Totowa, NJ: Little, Adams.

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

Prinz, W. (this issue) Import theory: The social making of consciousness, *Journal of Consciousness Studies*, **26** (3–4).

Rabinowitz, N.C., Perbet, F., Song, F., Zhang, C., Ali Eslami, S.M. & Botvinick, M. (2017) Machine theory of mind, *Computer Science arXiv*, [Online], 1802.007740.

Rosenthal, D. (2006) *Consciousness and Mind*, Oxford: Oxford University Press.

Saxe, R. & Kanwisher, N. (2003) People thinking about thinking people: fMRI investigations of theory of mind, *NeuroImage*, **19**, pp. 1835–1842.

Saxe, R. & Houlihan, S.D. (2017) Formalizing emotion concepts within a Bayesian model of theory of mind, *Current Opinion in Psychology*, **17**, pp. 15–21.

Titchner, E.B. (1898) The feeling of being stared at, *Science*, **8**, pp. 895–897.

Webb, T.W. & Graziano, M.S.A. (2015) The attention schema theory: A mechanistic account of subjective awareness, *Frontiers in Psychology*, **6**, art. 500.

Webb, T.W., Igelström, K., Schurger, A. & Graziano, M.S.A. (2016) Cortical networks involved in visual awareness independently of visual attention, *Proceedings of the National Academy of Sciences USA*, **113**, pp. 13923–13928.

Webb, T.W., Kean, H.H. & Graziano, M.S.A. (2016) Effects of awareness on the control of attention, *Journal of Cognitive Neuroscience*, **2**, pp. 1–10.

Wimmer, H. & Perner, J. (1983) Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception, *Cognition*, **13**, pp. 103–128.

Winer, G.A., Cottrell, J.E. & Karefilaki, K.D. (1996) Images, words and questions: Variables that influence beliefs about vision in children and adults, *Journal of Experimental Child Psychology*, **63**, pp. 499–525.

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**, pp. 417–424.

Yoshida, W., Dolan, R.J. & Friston, K.J. (2008) Game theory of mind, *PLoS Computational Biology*, **4**, pp. 1–14.