Part Three

Metaphilosophy of Consciousness Studies

11

Understanding Consciousness by Building It

Michael Graziano and Taylor W. Webb

1 Introduction

In this chapter we consider how to build a machine that has subjective awareness. The design is based on the recently proposed attention schema theory (Graziano 2013, 2014; Graziano and Kastner 2011; Graziano and Webb 2014; Kelly et al. 2014; Webb and Graziano 2015; Webb, Kean and Graziano 2016). This hypothetical building project serves as a way to introduce the theory in a step-by-step manner and contrast it with other brain-based theories of consciousness. At the same time, this chapter is more than a thought experiment. We suggest that the machine could actually be built and we encourage artificial intelligence experts to try.

Figure 11.1 frames the challenge. The machine has eyes that take in visual input (an apple in this example) and pass information to a computer brain. Our task is to build the machine such that it has a subjective visual awareness of the apple in the same sense that humans describe subjective visual awareness. Exactly what is meant by subjective awareness is not *a priori* clear. Most people have an intuitive notion that is probably not easily put into words. One goal of this building project is to see if a clearer definition of subjective awareness emerges from the constrained process of trying to build it. Our constraint is severe: Whatever we build into the robot must be possible given today's technology. Not every detail need be specified. This chapter discusses general concepts and will not come anywhere near a wiring diagram. Each component must nevertheless be something that, in principle, could be built.

2 Objective awareness

We start by giving the machine information about the apple, as depicted in Figure 11.2. In the case of a human, light enters the eye and is transduced

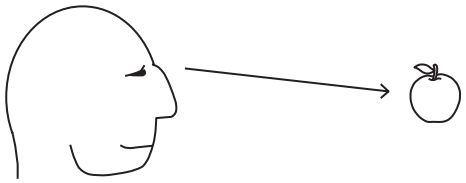

Figure 11.1 A robot has eyes looking at an apple.

into neuronal signals. Those signals are processed in the visual system, which constructs information that describes features of the apple. Those features include overall shape, local contour, colour, size, location and many other attributes bound together to form what is sometimes called an internal model. The internal model of the apple is a packet of information that is constantly updated as new information arrives from the eyes.

One of the most consequential properties of the internal model is its inaccuracy. It is an approximation or sketch. Borders are exaggerated, blurred visual features are partly filled in by algorithms that compute what is likely to be present, and at least one aspect of the apple, colour, has surprisingly little correspondence to the real world. The visual system does not contain information about wavelength that a physicist might want. It does not construct information about electromagnetic radiation, a continuous spectrum, absorption and reflection or transmission of light to the eye. The brain does not encode the actual reflectance spectrum of the apple. Instead it uses heuristics to compute a simplified property, colour and assign it to a spatial location, the surface of the apple. The reason why the brain's internal models are incomplete, one might even say cut-corner, is presumably for speed and efficiency in the face of limited resources. The brain must construct thousands of internal models and update them on a sub-second timescale.

With a camera and a computer we can give our robot just such a simplified internal model of the apple, as illustrated in Figure 11.2.

Is the robot in Figure 11.2 aware of the apple? In one sense, yes. Figure 11.2 illustrates what is sometimes termed objective awareness (Szczepanowski and Pessoa 2007). Information about the apple has gotten in and is being processed. That visual information could be used to drive behaviour. The machine is objectively aware of the apple in the same sense that a laptop is objectively aware of the information typed into it. But is the robot *subjectively* aware of the apple? Does it have a conscious visual experience? At least some scholars suggest that subjective awareness emerges naturally from information processing (Chalmers 1997). In that view even a thermostat is subjectively conscious of the simple

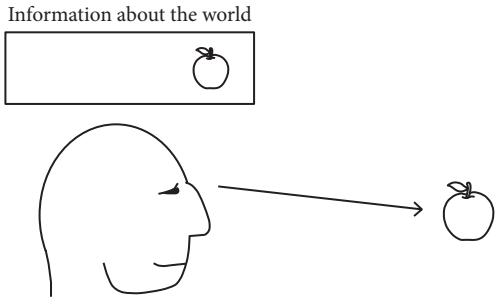

Figure 11.2 The robot has information about the world in the form of internal models. Here the robot has been given an internal model of the apple. The robot is objectively aware of the apple. We suggest this is not a complete account of subjective awareness.

information that it processes. By extension, visual subjective awareness arises from visual processing and therefore our machine is subjectively aware of the visual stimulus. The hypothesis, however, is fundamentally untestable. If that hypothesis is correct then we are done building an aware machine. Anything that manipulates information – which may well be everything in the universe – is conscious of the information it manipulates. And we will never be able to confirm the proposition. This approach is sometimes called 'panpsychism' (Skrbina 2005). Rather than stop with this non-testable answer, in the following sections we continue our exploration to see if we can gain a clearer insight into subjective awareness.

3 Cognitive access

In Figure 11.3, a new piece has been added to the machine. To be able to query the machine, we have added a user interface. It is a search engine, a linguistic/cognitive component. We can ask it a question. It searches the database of the internal model, and on the basis of that information answers the question. Again, everything in Figure 11.3 is in principle buildable with modern technology.

We ask the machine, 'What's there?' The search engine accesses the internal model, obtains the relevant information and answers, 'An apple'. We ask, 'What are the properties of the apple?' The machine answers, 'It's red, it's round, it has a dip at the top, it has a stem protruding upward from the dip', and so on.

The robot in Figure 11.3 could represent an entire category of theory about consciousness. In it, higher-order cognition has access to a lower-order sensory representation. For example, in the global workspace theory (Baars 1988), the

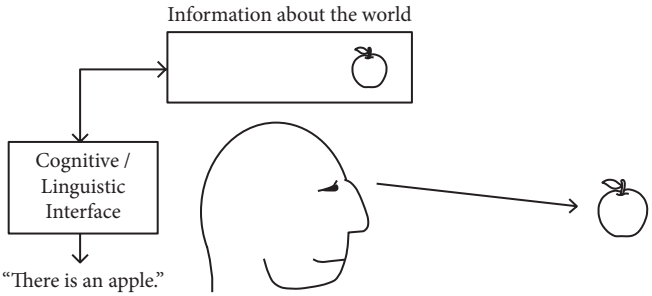

Figure 11.3 The robot has a linguistic interface that acts as a search engine. It takes in questions from the outside, searches the internal model and on the basis of that information replies to the question. The robot has a type of higher cognitive layer that can access a lower-order sensory representation. Yet we suggest this is still an incomplete account of subjective awareness.

information in the sensory representation is broadcast to other systems in the brain, allowing higher cognition to gain access to it. As a result, we can ask the machine about the apple and it can answer. We suggest, however, that Figure 11.3 represents an incomplete account of consciousness. To highlight that incompleteness, we ask the machine, 'Are you aware of the apple?' The search engine accesses the internal model and obtains no answer to the question. The internal model does not contain any information about the property of awareness. It does not even have information about the item 'you'. Of the three key words in the question, 'you', 'aware', and 'apple', the search engine can return information only on the third. Equipped with the components shown in Figure 11.3, the machine cannot even compute in the correct domain to answer the question. You might as well ask your digital camera whether it is aware of the pictures it takes. It cannot process the question.

4 Self-knowledge

If the difficulty with the machine in Figure 11.3 is that it lacks sufficient information, we can try to fix the problem by adding more information. In Figure 11.4 we add a second internal model, a model of the self. In the human brain, the self-model is complex and probably spans many brain regions. It is probably more accurately described as a collection of many models. It might include information about the shape and structure and movement of the physical

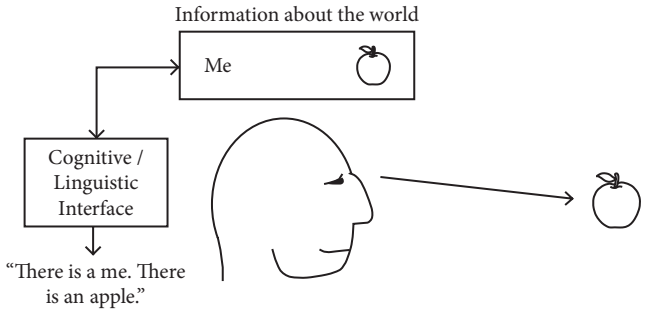

Figure 11.4 The robot has a second internal model, a model of the self. The self-model may contain information about the physical body and how it moves, autobiographical memory and other self-information. The robot now has self-knowledge. Yet we suggest this is still an incomplete account of subjective awareness.

body, autobiographical memory and information about one's personality and behavioural habits.

We ask the machine in Figure 11.4, 'Tell us about yourself'. Unlike in the last iteration, this time the machine can answer. It has been given the construct of self. It might say, 'I'm a person. I'm this tall, this wide, I can move my arms and legs, I've got brown hair, I like Beethoven, I'm friendly' and so on. It can provide information from a rich internal model of self.

Again, Figure 11.4 could represent an entire category of theory about consciousness. Many theories relate consciousness to self-information or self-narrative (Gazzaniga 1970; Nisbett and Wilson 1977). Self-knowledge is clearly an important part of what many people consider to be consciousness. But once again, we suggest that as a theory of consciousness, Figure 11.4 is incomplete. To make the point, we ask the machine another question: 'What is the mental relationship between you and the apple?' The search engine searches the internal models. It obtains information about the self, separate information about the apple, but no information about the mental relationship between them. It has no information about what a mental relationship even is. Equipped only with the components in Figure 11.4, it cannot answer the question.

5 The attention schema

The machine in Figure 11.4 has an internal model of the apple and an internal model of the self, but it lacks an internal model of a third crucial part to the scene: the computational relationship between the self and the apple. The machine is

focusing its processing resources on the apple. It is attending to the apple. To try to improve the machine, we add one more internal model, a model of attention.

First we clarify what we mean by attention, given that the term has been used in so many different contexts. By attention we refer to an entirely mechanistic process that can in principle be duplicated with modern technology. Our use of the term is based on a neuroscientific theory, the biased competition theory (Desimone and Duncan 1995; Beck and Kastner 2009). In that theory, signals in the brain compete with each other due to lateral inhibitory processes. One or a small number of signals may temporarily win the competition, momentarily rising in signal strength while suppressing other signals. The winner of the competition, due to its greater signal strength, has an exaggerated influence on other systems in the brain such as memory and response selection. That competition among signals can be biased towards one or another winning signal by a variety of modulating signals. Attention, in this mechanistic account, is a process by which the brain focuses computing resources on a limited set of signals.

Consider the case of a person attending to an apple. The apple is probably only one of many items in what may be a cluttered visual scene. The internal model of the apple has won the competition of the moment and other visual models are relatively suppressed. As a result, the apple's internal model can dominate the brain's outputs. Since this process of attention is mechanistic and in principle buildable, given current technology, we are allowed to add it to our robot. There are now three fundamental components to the scene: an apple, a self and an attentive relationship between them.

The machine in Figure 11.4 contains an internal model of the self and of the apple, but has no internal model of attention. What would happen if we added an internal model of the machine's attentional relationship to the apple? The internal model of attention, like all internal models, is information. It is a continuously updated set of information. It does not present any fundamental engineering problem. It is in principle buildable. We are allowed to add it to our robot. Figure 11.5 shows this addition to the machine.

We first consider what information would be included in an internal model of attention. Attention has a complex neuronal mechanism, complex dynamics and complex consequences. But not all of the microscopic details of attention need be represented in the internal model. Like the internal model of the apple, the internal model of attention would presumably describe useful, functional, abstracted properties. We should not expect an internal model of attention to be a scientifically accurate description of attention. As illustrated in Figure 11.5, an internal model of attention might describe attention as a mental possession of

something. It might describe attention as something that empowers the machine to react to the attended item. It might describe attention as something located inside oneself. These are only three general, abstracted properties of attention. An internal model of attention might include a great deal more information about attention, about its consequences and dynamics. But an internal model of attention would not contain information about neurons, synapses, lateral inhibitory processes, competition among electrochemical signals and other microscopic aspects of attention. The internal model would be silent on the physical mechanism of attention. In the same way, the internal model of the apple assigns a colour to the apple's surface while leaving out the physical details of electromagnetic waves. The internal model of attention would be as incomplete and inaccurate as any other internal model. We term this internal model of attention the 'attention schema' in parallel to the internal model of the physical body, the 'body schema' (Graziano and Botvinick 2002).

Now that we have added an internal model of attention, we ask the machine in Figure 11.5 more questions. We ask, 'What object are you looking at?' It can still answer. The cognitive/linguistic search engine can still find that information among the internal models. The machine says, 'An apple'. We ask, 'Tell us about yourself'. It can answer this just as before. 'I'm a person'. Now we ask, 'What is the mental relationship between you and the apple?' This time, the search engine can

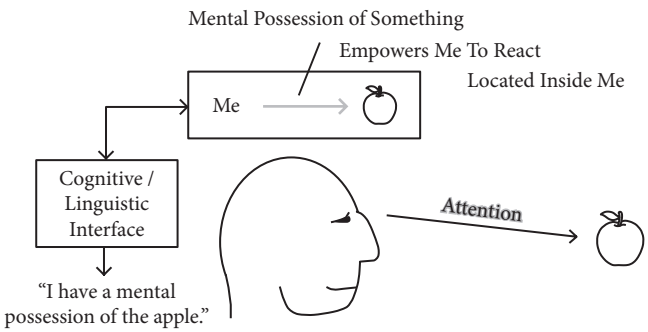

Figure 11.5 The main components of the attention schema theory. The robot has an internal model of the self, an internal model of the apple, and a third internal model, a model of the attentional relationship between the self and the apple. Attention here refers to the brain focusing its processing resources on the apple. The internal model of attention describes that computational relationship in an abstracted, schematic manner. The attention schema is accurate enough to be useful but not so accurate or detailed as to waste resources and processing time. The attention schema describes something impossible and physically incoherent, a caricature of attention, subjective awareness. This machine insists that it has subjective awareness because it is captive to the incomplete information in its internal models.

return an answer. Reporting the information obtained from its internal models, the machine answers, 'I have a mental possession of the apple'.

We ask the machine for more details of this mental possession. First, however, we ask a question of clarification. We ask, 'Do you know what is meant by the physical properties of something?' The machine has a self-model that includes a body schema, a description of the physical self. It also has an internal model of the apple that describes a physical object. Therefore it can answer, 'Yes, I know what physical properties are'. We then ask, 'What are the physical properties of this mental possession?'

The machine in Figure 11.5 reports the available information. It says (and here we are guilty of giving it a sophisticated verbal capability), 'My mental possession of the apple, the mental possession itself, has no describable physical properties. Yet it exists. It is a part of me. It is inside me. It is my mental possession of things. It enables me to react to things. There is a me, there is an apple, but that is not a complete description. There is something else, something with no physical substance and yet a spatial location inside me, something metaphysical, the mental relationship between subject and object. I have a subjective mental experience of the object. Likewise, I have a subjective mental experience of each component of the object – of its color, of its shape, of its size'.

The machine is claiming to have subjective awareness.

We should not be surprised by this response. We know why the machine behaves the way it does because we built it. It accesses internal models and whatever information is contained in those models it reports to be true. It reports a physically incoherent property, a metaphysical essence of consciousness, because its internal models are incomplete descriptions of physical reality.

The machine, however, does not know why it answers the way it does. It has no information about how it was built. Its knowledge is limited to the contents of its internal models, and those models do not contain the information, 'By the way, this is an internal model composed of information that is incomplete and sometimes wrong'.

To try to probe the limits of the machine's responses, we ask, 'Are you just a machine accessing the information in internal models, and is that why you claim to have subjective awareness of the apple?' The machine accesses its internal models and, based on the information found there, answers, 'I don't know what internal models are. I don't have them. There is an apple, there is a me, and I am aware of the apple'.

The machine is trapped in the same ego tunnel described by Metzinger (2010). It is capable of introspection, in the sense of cognitive machinery accessing deeper internal information. It is capable of self-knowledge. It is capable of knowledge about its relationship to the world and to itself. But it is captive to the incomplete information in its internal models. Introspection will always return the same answer. It insists it has subjective awareness because, when its internal models are searched, they return that information.

It should be possible, in principle, to build a brain that insists on anything we like. We just need to insert the right packet of information into its low-level internal models. We could build a brain that insists it is a cosmic squirrel instead of a brain. We could build a brain (as evolution did) that insists that white light is pure brightness scrubbed clean of all contaminating colours. The attention schema in Figure 11.5 is the right packet of information to lead a brain to conclude and report and insist that it has a hard problem – a non-physically describable subjective awareness of the apple.

There is no reason to limit the machine to visual awareness of an apple. The same logic could apply to anything. Along similar lines, we could build the machine to be aware of a touch; aware of a sound; aware of its own body; aware of memories from its past that are being recalled and replayed; aware of the thought that it is aware. If the machine can direct mechanistic attention to signal X, and if the machine includes an internal model of that attentional state, then the machine is equipped to insist, 'I am aware of X'.

6 Possible misconception: Higher cognition and the attention schema

One possible misconception about attention schema theory is that it depends on higher-order cognition, as in the higher-order thought theory (Lau and Rosenthal 2011). It does not. The cognitive/linguistic layer was added as a convenience to be able to query the machine. We could remove it. We could build something more like a rat than a person, something with little cognitive ability and no language ability. Imagine crossing out the cognitive/linguistic layer in the machine in Figure 11.5. The machine would still have its internal models. The internal models are fundamental, low-level representations of the world. They are useful for survival. They are the brain's most fundamental simulation of the world, and in that simulation, you are conscious of the apple. That simulation can be constructed even if the brain in question lacks the

capability to introspect, to cogitate or to talk. Attention schema theory is in this sense unrelated to higher-order thought, although the cognitive layer makes a convenient interface for talking to the machine.

7 Possible misconception: What generates actual awareness?

In a naïve, intuitive approach to the topic, it is natural to think that subjective awareness is a non-physically describable, private essence inside us. One can then ask what is the specific brain process that, when run, causes awareness to emerge? One could mistakenly think that we are putting forth the attention schema as our candidate for that special device that, when operated, generates awareness. This mistaken line of thinking leads to a question: Why would an internal model of attention generate subjective awareness? Accepting that we probably have an internal model of attention – which seems reasonable – why would such a thing generate an inner feeling? Where is the logic in that? We hit the same obstacle as always in understanding consciousness. How can science cross the gap from physical mechanism to metaphysical awareness?

The answer is simple. In attention schema theory, nothing generates awareness. Awareness is not generated. The brain constructs information that describes attention. The information is neither complete nor entirely accurate. What is described, instead, is a physically impossible thing, a spooky caricature of attention – subjective awareness. The brain is captive to its own internal information. The entire world known to the brain is defined by the incomplete information it constructs.

As an analogy, consider the mystery of white light. Before Newton, scholars asked a naïve question: What washes light so that it becomes clean and uncontaminated by colours? Likewise, what dirties white light, contaminating it to become coloured light? These highly intuitive misconceptions are the result of an internal model constructed deep in the early layers of the visual system. That internal model evolved over millions of years presumably because it is a useful, though incomplete and imperfect, way to model some aspects of light. The question of how white light becomes purified is truly a hard problem of light. When we mistakenly take our internal models to be literally accurate, we run into unsolvable scientific hard problems. Here, of course, 'hard problem' is a euphemism for 'ill-posed problem'.

8 The parable of the Heliocentric theory

Ptolemy and Galileo walk into a bar. They strike up a conversation.

Ptolomy: Your theory is silly. I spotted the error right away. It doesn't solve the hard problem. What pushes the sun around the stationary earth? I think it may be the chariot of Helios. In your theory, the movement of the sun around the earth is left entirely unexplained.

Galileo: No, the theory is explanatory. You see, the earth orbits the sun.

Ptolomy: I still don't see how that explains the motion of the sun around the earth.

Galileo: It doesn't. The sun doesn't move around the earth.

Ptolomy: Ah ha! Now I know how to pigeon-hole your theory. You 'solve' the hard problem by denying the phenomenon exists in the first place. But that's a cop-out. We're both philosophers, so let's be more systematic in our logic. The definition of motion is: it moves. Now look at the earth: it ain't moving. QED.

Galileo: But the theory explains why you think that. You see, the principle of Galilean relativity means that in the closed system of you and the earth, there is no observation you can make to show that the earth is moving. It's not possible. That's why all your observations tell you that the earth is stationary. You can't know the answer unless you look outside that closed system. If you study the stars and planets and sun, and take a hypothetical external perspective, you can infer the truth. But the scientific truth will always differ from your immediate, limited, earth-bound observations.

Ptolomy: That sounds complicated. Bottom line: you've failed to explain how the sun moves around the earth. The hard problem remains unanswered. Indeed, you've left it unaddressed. Therefore, friend, I'm afraid I must reject your theory. But since I won the argument, I'll be generous and buy you a drink.

9 Uses of the attention schema

It seems obvious why the brain would evolve a visual system capable of constructing models of visual objects. It allows the animal to navigate in its visual environment. It also seems obvious why the brain would evolve an elaborate self-model, especially a model of the physical body. Monitoring and predicting the changing state of your own body is useful in controlling movement. The adaptive usefulness of an attention schema is less obvious but, as we argue below, much more profound.

One of the traps in evolutionary thinking is to suppose that a trait has one function or was shaped by only one type of evolutionary pressure. Even when the evolutionary 'purpose' of a trait seems obvious – for example, teeth are obviously adapted for chewing – other functional roles can turn up unexpectedly. After all, teeth are also partly adapted for social signalling. It is possible that an internal model of attention serves many adaptive roles, and perhaps even has a different mixture of functions in different species. We suggest at least three major adaptive functions but of course there may be others. These three proposed functions are described in the following sections.

10 The attention schema is useful for the control of attention

In control theory, if you want to build a capable control system, you should give it an internal model of the thing to be controlled (Camacho and Bordons Alba 2004). For example, the brain constructs a body schema, an internal model of the body, to help control movement (Graziano and Botvinick 2002; Scheidt et al. 2005; Wolpert, Goodbody and Husain 1998). We suggest that one adaptive function of an attention schema is to help control attention. We also suggest that this function may be the evolutionary origin of awareness.

Attention is probably evolutionarily old. Some aspects of selective signal enhancement can be seen in insects, crabs, birds and mammals (Barlow and Fraioli 1978; Beck and Kastner 2009; Mysore and Knudsen 2013; van Swinderen 2012), which shared a common ancestor more than half a billion years ago. There is little use in having attention without any ability to control it. Therefore we suggest that at least half a billion years ago nervous systems began to evolve a dynamical systems controller to regulate attention, and one part of that control system was an internal model of attention. In this proposal, the rudiments of consciousness are extremely ancient and widespread in the animal kingdom. Some form of attention schema, at least a simple internal model of attention used to help control attention, could be present in the brains of almost all animals that have brains. We would probably not recognize a simple internal model of attention as similar to our human awareness. Over the intervening millions of years, the attention schema may have evolved the rich information that we recognize as subjective awareness.

No internal model in the brain is perfect. For example, the body schema makes errors, becoming misaligned from the body, incorrectly representing the movement and configuration of the body. In those cases, the control of the

body is impaired. When the internal model of your arm is temporarily off, you make errors in moving your arm. Like all internal models, the attention schema should also sometimes make errors. When those errors occur – when awareness becomes misaligned from attention – then the control of attention should suffer in specific ways predictable from dynamical systems control.

To clarify how attention and awareness can separate, consider Figure 11.5 again. The machine directs visual attention to an apple. An internal model of that attention is constructed in the machine. That internal model might sometimes make errors. To be clear, we are not talking about an error in the internal model of the apple, which might lead to a visual illusion. We are also not talking about an error in the internal model of the self, which might lead to a body image distortion and inaccurate movement control. We are talking about an error in the internal model of attention itself. There may be many possible kinds of error, all worth exploring theoretically and experimentally. Here, we focus on one particularly simple kind of mismatch between awareness and attention that is convenient to approach in a practical experiment: attention in the absence of awareness.

In the past decade, many experiments have demonstrated that people can attend to a visual stimulus in the absence of awareness of that stimulus (Hsieh, Colas and Kanwisher 2011; Jiang et al. 2006; Kentridge, Nijboer and Heywood 2008; Koch and Tsuchiya 2007; Lamme 2004; McCormick 1997; Norman, Heywood and Kentridge 2013; Tsushima, Sasaki and Watanabe 2006). These experiments almost always involve a dim visual stimulus or a visual stimulus that is masked by a second visual display, putting the stimulus at the edge of detectability. Even when people assert that they see no visual stimulus, it can still draw their attention, improving the processing of subsequent stimuli at the attended location. The significance of attention in the absence of awareness has been debated, but attention schema theory provides a simple explanation. In the theory, awareness is the internal model of attention. Attention without awareness occurs when the internal model fails to update correctly.

Because control theory is a well-developed area of research, it is possible to put attention schema theory to experimental test. Measure attention to a visual stimulus. Manipulate the visual display such that in one condition, the participants report being subjectively aware of the stimulus, and in another condition, the participants report being subjectively unaware of the stimulus. Compare the aware and the unaware conditions. Three results should be obtained. First, attention should still be possible without awareness. Second, attention should change without awareness. Third, the changes should match

the pattern predicted by control theory. Without an internal model, the control of attention should suffer in specific ways.

To conduct these tests, we used a Posner paradigm (Posner 1980), a standard method for measuring visual attention in human participants. The logic of the Posner paradigm is easily explained. The participant looks at the centre of a computer screen. A small dot is briefly flashed to the right or left. After the dot is gone, within a fraction of a second, another stimulus, the test stimulus, is presented to the right or left. The participant's task is to discriminate the test stimulus as quickly as possible and respond by pressing a key on a keyboard. For example, in some tests the stimulus might be an A or an F and the participant must press the A or F key as quickly as possible. If the test stimulus appears at the same location as the initial dot, the reaction time to the test stimulus is generally fast. This occurs because the dot initially and automatically draws attention to the correct location. If the test stimulus appears on the opposite side of the screen as the dot, the reaction time to the test stimulus is usually slower by a few tens of milliseconds. This occurs because the dot initially draws attention to the wrong location. By measuring the difference in reaction time between these two conditions, it is possible to infer how much visual attention was drawn to the initial dot. The key to the experiment is that the participant does not need to be aware of the dot. If the dot is dim, extremely brief or masked by other visual stimuli, the participant may claim never to see it. Even so, the dot can still draw attention and thereby affect the response time to the subsequent test stimulus. In this way, it is possible to measure attention to the dot whether or not the participant is aware of it. To determine whether the participant was subjectively aware of the dot, at the end of each trial the participant is asked whether the dot was seen.

This method allows us to test how attention to a visual stimulus changes when people are no longer aware of the dot. If awareness serves as the internal control model of attention, then predictable changes should occur. For example, without an internal model, attention should become less stable. The controller no longer has information about the current state of attention and therefore can no longer adjust in real time to compensate for fluctuations – like balancing a stick on your hand with your eyes closed and therefore no information about what the stick is doing. In addition to a loss of intrinsic stability, attention should also become more sensitive to outside influences such as the brightness of the stimulus. Again, it is like balancing a stick on your hand with your eyes closed, this time while the stick it being nudged by someone else. It is harder to compensate for the nudge.

Figure 11.6 shows data from one experiment (Webb, Kean and Graziano 2016). The Y axis shows the amount of attention drawn to the dot. The X axis

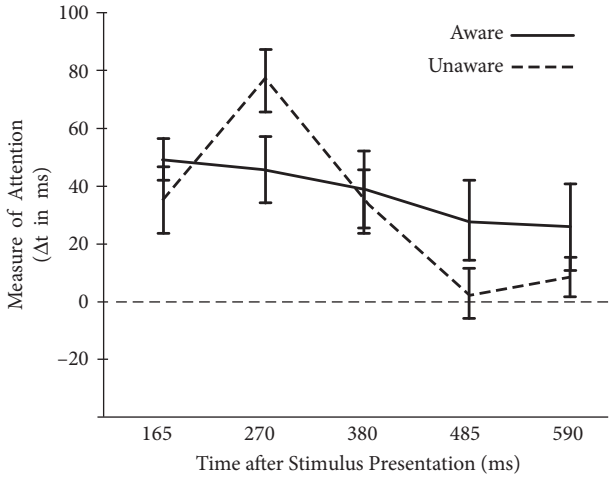

Figure 11.6 Testing attention with and without awareness. In this experiment, attention to a visual stimulus is tested by using the stimulus as a cue in a Posner spatial attention paradigm (see Webb, Kean and Graziano 2016 for details). In some trials, the participants are aware of the visual cue (dotted line). In other trials, they are unaware of it (bold line). Attention to the cue is less stable across time when awareness is absent. This result follows the predictions of control theory in which an internal control model helps to maintain stability of the controlled variable. The X axis shows time after cue onset. The Y axis shows attention drawn to the cue (Δt = [mean response time for spatially mismatching trials in which the test target appeared on the opposite side as the initial cue] – [mean response time for spatially matching trials in which the test target appeared on the same side as the initial cue]). Error bars are standard error.

shows the time at which attention was measured relative to the onset of the dot. The dotted line shows the results when participants were subjectively aware of the dot. A significant amount of attention was drawn to the dot and that attention slowly decreased over time. The bold line shows the results when participants were not subjectively aware of the dot. Here attention was still drawn to the dot, yet the time course of attention changed in the absence of awareness. At one time point, participants actually paid significantly more attention to the dot when they were unaware of it. At another time point, they paid less attention when they were unaware of it. Without awareness, attention was less stable over time.

Figure 11.7 (Webb, Kean and Graziano 2016) shows data from another, similar experiment. When participants were aware of the dot, they paid slightly more attention to a brighter dot than to a dimmer dot, as might be expected. When participants were not aware of the dot, attention was much more sensitive to the brightness of the stimulus.

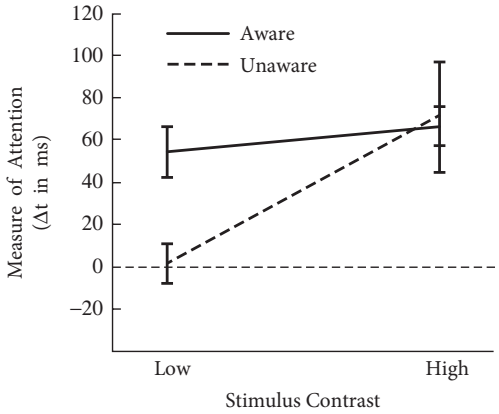

Figure 11.7 Testing bottom-up attention with and without awareness. In this experiment, bottom-up attention to a visual stimulus is tested by using the stimulus as a cue in a Posner spatial attention paradigm (see Webb, Kean and Graziano 2016 for details). In some trials, the participants are aware of the visual cue (dotted line). In other trials, they are unaware of it (bold line). Attention to the cue was more sensitive to the visual contrast of the cue when awareness was absent. This result follows the predictions of control theory in which an internal control model helps to resist perturbations. The Y axis shows attention drawn to the cue (Δt = [mean response time for spatially mismatching trials in which the test target appeared on the opposite side as the initial cue] – [mean response time for spatially matching trials in which the test target appeared on the same side as the initial cue]). Error bars are standard error.

These and other experiments (Webb and Graziano 2015; Webb, Kean and Graziano 2016) tell us that awareness is not an epiphenomenon. It actually does something. It is important for the mechanistic control of attention. When people are not aware of something, they can still pay attention to it, they can even pay approximately the same amount of attention to it, and in some circumstances may even pay more attention to it. But the control of attention changes. Attention is less stable in time and more easily perturbed by external influences. These two specific deficits are predicted from the loss of an internal model. Experiments like these point to a specific relationship between attention and awareness, that awareness acts like the internal model of attention. Without awareness, attention is still possible but the control of attention suffers in predictable ways. Even fast, automatic aspects of attention, within the first few hundred milliseconds of stimulus onset, depend on that internal model of attention. The data in Figure 11.7 show attention only fifty milliseconds after stimulus onset. In fifty milliseconds there is no time for high-level cognition, volition or choice, and yet the control of attention still depends on the presence of subjective awareness.

This type of result suggests that awareness serves a specific mechanistic function as the internal model of attention.

11 The attention schema is useful for the integration of information

A common hypothesis is that consciousness is related to the widespread integration of information in the brain. One of the earliest examples is the global workspace theory first proposed by Baars (1988), in which many disparate types of information are pooled together by attention into a single whole that allows for more intelligent, coherent guidance of behaviour. Another similar proposal is that consciousness is caused by binding information together (Crick and Koch 1990). A more recent proposal is that the amount of integrated information can be mathematically quantified and a large amount of integrated information is associated with subjective experience (Tononi 2008). Other variants of the global workspace and integration-of-information hypothesis have been proposed (Dehaene 2014).

One difficulty with this category of theory, at least as it is usually presented, is the metaphysical gap. Granted that disparate information in the brain is integrated, why would that cause subjective awareness of any of the information in question? The integrated information approach suffers from an intuitive bias, the notion that subjective awareness is a metaphysical thing that is generated by some process in the brain. If we can figure out what physical process generates awareness, then we are as close as we can ever be to explaining awareness, while still leaving unexplained the gap from physical mechanism to metaphysical essence. That is the conceptual approach of almost all theories of consciousness.

Attention schema theory adds a missing piece to the account of integrated information. It adds the attention schema, a chunk of information that describes subjective awareness.

In Figure 11.5, three internal models are diagrammed: an internal model of the self, of the apple and of the attentional focus of the self on the apple. These internal models are information constructed in the brain. Collectively they form one larger, integrated model describing how the self is aware of the apple. The internal model of attention, the attention schema, acts as a connector. Without that intermediate piece, the brain would have two disconnected internal models. Like the brain in Figure 11.4, it could say, 'There is a me', and separately, 'There is

an apple'. The attention schema bridges between information about the self and information about the world – not just the external world, but sometimes also aspects of the internal world to which attention is directed.

The attention schema, if it is to actually model attention, must be a universal connector. It must link to any type of information to which the brain can attend. You can attend to a touch, to a sound, to a recalled memory, to a specific thought, to an emotion. Therefore, to model the state of attention, the attention schema must be able to connect to all of those information domains. By its nature, an attention schema must serve as an integrative hub. The attention schema is a chunk of information that is uniquely connectable to many other chunks of information.

In many ways the attention schema theory shares features with the global workspace theory and integrated information theory. The difference is that it explains the consciousness part. It avoids the metaphysical gap. It does not postulate that pooling information, by itself, unexplainably generates metaphysical awareness. Instead it adds to that pool of information a specific ingredient that is easy to overlook but is of paramount importance. It adds the attention schema, a chunk of information that describes metaphysical awareness and causes us to assert that we have awareness.

12 The attention schema is useful for social perception

We first arrived at the attention schema theory by considering social perception (Graziano 2010; Graziano and Kastner 2011). In a social context people attribute mind states to each other, a process sometimes called theory of mind – the ability to construct theories of other people's minds (Frith and Frith 2003; Wimmer and Perner 1983). People attribute beliefs, emotions, intentions and other mind states to each other. By 'attributing' mind states, what is meant is not an intellectual, cognitive process of deducing what is likely to be in other people's minds. Instead social attribution is automatic, intuitive and can even contradict what we know intellectually. For example, people can have a powerful illusion of mind states in a puppet while knowing intellectually that the puppet has no mind. Beneath the level of higher cognition, the human brain constructs internal models of other people's minds.

We suggest that awareness is one of the most basic mind states that people attribute to each other. It is difficult to attribute to John an intention to reach for a nearby coffee cup unless you can intuitively understand that John is aware of the coffee cup. It is difficult to attribute to John any anger towards the vandal who

damaged his car, unless you can understand that John is aware of the damage. If you think he is not aware of the damage, you would suppose he is not angry.

Suppose you are watching John. John is directing attention to a coffee cup, in the mechanistic sense of attention in which signals in the brain compete with each other and the signals related to the coffee cup win that competition. Those signals impact widespread systems in his brain and dominate his behaviour. He is able to process the coffee cup, reach for it, avoid knocking it over or remember it for later. Nothing determines his immediate and future behaviour as much as his state of attention. If you want to predict John's behaviour, as a first pass computation it would be of the utmost importance to have an attention schema that can model John's state of attention. That attention schema would not reconstruct the mechanistic details of John's attention. Your brain has no access to the details inside John's brain. Your brain has no use for an internal model of neuronal dynamics inside John's brain. Instead, that attention schema would model a simpler property stripped of mechanistic details. John has a mental possession of the apple and that mental possession has certain basic dynamics and consequences for behaviour. Your brain would attribute the property of *awareness* to John, as though it were a metaphysical essence inside him, because awareness is a good heuristic model of attention. It was from this consideration of the social use of awareness that the more general theory emerged in which awareness, whether attributed to someone else or to oneself, is a model of the attentional process (Graziano 2010; Graziano and Kastner 2011).

Exactly when or how the social use of awareness evolved is not clear. Apes have a well-developed theory of mind (Premack and Woodruff 1978; Wimmer and Perner 1983), but social attribution of awareness may have predated primate evolution by many millions of years. Dogs show an ability to intuit the attentional state of other dogs (Horowitz 2009). Crows have some elements of a theory of mind (Clayton 2015). Reptiles show highly complex social behaviours (Brattstrom 1974). Since birds, reptiles and mammals diverged at least 300 million years ago in the Carboniferous period, the ability to model the attentional state of others may have emerged early, though it presumably is better developed in some species than others.

We suggest that an attention schema first evolved as part of the mechanism for controlling attention as discussed in previous sections. It gradually expanded its role to become a central mechanism for the integration of information across disparate domains, leading to more flexible and intelligent behaviour. A third main function also gradually emerged, modelling the attentional states of other animals to help predict their behaviour. This social function of the attention

schema may be present in some form in a large range of animals. In humans the ability is especially well developed. We are so prone to attribute awareness that we live immersed in a world painted with projected consciousness. Human spiritual belief is arguably a manifestation of an exuberant social machinery.

If the brain constructs an attention schema and uses it to model others, as well as oneself, then perhaps overlapping brain areas are involved in attributing awareness to others and oneself. At least some evidence suggests that this is the case in humans. One brain region where these two functions may overlap is the temporoparietal junction (TPJ). Clinical evidence shows that damage to the TPJ can cause a severe neglect syndrome in which patients are unaware of anything in the half of space opposite the lesion (Critchley 1953; Halligan et al. 2003; Vallar 2001; Valler and Perani 1986). Typically, right brain damage leads to left spatial neglect, though the pattern can also sometimes reverse. In neglect, patients can still process sensory information from the affected side and can sometimes react to it, but claim a lack of awareness of it. Yet the TPJ also plays a role in attributing mind states to others. When brain activity is measured in an MRI scanner while people engage in social perceptual tasks, the TPJ is consistently active (Saxe and Kanwisher 2003; Young, Dodell-Feder and Saxe 2010). These lines of evidence show that two seemingly unrelated functions, social perception and the construction of one's own awareness, are at least in close proximity in the brain.

To determine just how much overlap there may be between these two functions in the TPJ, we performed an experiment (Kelly et al. 2014). We measured the brain activity of participants in an MRI scanner while they engaged in a social cognition task, rating whether a cartoon character was aware of an object next to it. Elevated activity associated with this task was found in a part of the TPJ. Each subject had a definable zone of activation, a hotspot. We then took the participants out of the scanner and tested the effect of disrupting that hotspot. The disruption depended on a technique called transcranial magnetic stimulation. In that technique, a magnetic pulse is passed through the skull and for a fraction of a second disrupts the neuronal activity in a small area of cortex approximately one centimetre in diameter. When the hotspot on the one side of the brain was disrupted, subjects were less able to detect visual stimuli on the opposite side of space. When a different part of the TPJ was disrupted, a part that did not become active in the social perception task, no effect was seen on participants' ability to detect stimuli.

To summarize this experiment, when people looked at a face and answered the question, 'Is he aware of the object next to him?', a specific hotspot in the brain became active. When that specific hotspot was disrupted, people were less able to be aware of objects next to themselves. The results suggest that the

networks in the brain that attribute awareness to others physically overlap the networks that construct one's own awareness.

13 Summary

The brain is an information-processing device. It takes in data, processes that data and uses it to help guide behaviour. When that machine ups and says, 'I have a magic essence inside me', rather than believing that literal proposition and then failing to obtain any scientific purchase on the magic, we can ask instead, 'How did the machine arrive at that quirky self-description? What is the utility of that self-description? What brain areas might be involved in computing that information?'

In attention schema theory, awareness is an impossible, physically incoherent property that does not exist and that is described by a packet of information in the brain. That packet of information is an internal model, and its function is to provide a continuously updated account of attention. It describes attention in a manner that is accurate enough to be useful but not accurate or detailed enough to waste time and resources. The brain is captive to the incomplete information in its internal models. To put it tautologically, the brain has no other information than the information it has. Hence people insist that they have subjective awareness, mystics wax poetic about it, philosophers and scientists dedicate themselves to understanding what subjective awareness is and how it is generated, and authors write chapters on the topic. In attention schema theory, awareness is not a total fabrication of the brain. It is not an illusion. It is perhaps best described as a caricature. It is a caricature of attention, a physical process that actually does exist and is of central importance in brain function. A shorthand way to describe the theory in five words is this: Awareness is an attention schema.

In the theory, the attention schema has at least three major adaptive uses. First, it is important in the control of attention. In dynamical systems theory, a good controller of attention should include an internal model of attention. Our data suggest that awareness does act as the internal control model of attention. We suggest that this function may be the evolutionary origin of awareness.

A second possible adaptive function of an attention schema is to promote the integration of information across disparate information domains. An attention schema by definition links information about the self with information about whatever is in the focus of attention. In this sense it serves as a connector of different types of information.

A third possible adaptive function of an attention schema is to promote social perception. If we use the attention schema to model the attentional states of others as well as of ourselves, in effect attributing awareness to others and to ourselves, then it could be foundational to social perception. We suggest that this social use represents a major evolutionary expansion of the attention schema and has reached a particularly elaborated state in humans.

The theory is extremely simple in concept and yet extremely difficult for many people to accept. The theory itself explains why people have such strong intuitions to the contrary. Introspection is cognitive machinery accessing internal models, and the internal model of attention informs us that we have a private, non-physically describable essence, a metaphysical property, a mental possession of things that empowers us to decide, to choose, to act, to remember. But the brain's evolutionarily built-in models are not accurate. They are caricatures of reality.

References

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

Barlow, R. B., Jr., Fraioli, A. J. (1978). 'Inhibition in the Limulus Lateral Eye *in Situ*', *Journal of General Physiology*, 71, 699–720.

Beck, D. M., Kastner, S. (2009). 'Top-Down and Bottom-Up Mechanisms in Biasing Competition in the Human Brain', *Vision Research*, 49, 1154–65.

Brattstrom, B. H. (1974). 'The Evolution of Reptilian Social Behavior', *American Zoology*, 14, 35–49.

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

Chalmers, D. (1997). *The Conscious Mind*, New York: Oxford University Press.

Clayton, N. S. (2015). 'Ways of Thinking: From Crows to Children and Back Again', *Quarterly Journal of Experimental Psychology*, 68, 209–41.

Crick, F, Koch, C. (1990). 'Toward a Neurobiological Theory of Consciousness', *Seminars in the Neurosciences*, 2, 263–75.

Critchley., M. (1953). *The Parietal Lobes*, London: Hafner Press.

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

Desimone, R., Duncan, J. (1995). 'Neural Mechanisms of Selective Visual Attention', *Annual Review of Neuroscience*, 18, 193–222.

Frith, U., Frith, C. D. (2003). 'Development and Neurophysiology of Mentalizing', *Philosophical Transactions of the Royal Society of London Biological Sciences*, 358, 459–73.

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

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

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

Graziano, M. S. A. (2014). 'Speculations on the Evolution of Awareness', *Journal of Cognitive Neuroscience*, 26, 1300–4.

Graziano, M. S. A., Botvinick, M. M. (2002). 'How the Brain Represents the Body: Insights from Neurophysiology and Psychology, in *Common Mechanisms in Perception and Action: Attention and Performance XIX*, 136–57 (ed.), R. Prinz and B. Hommel, 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.

Graziano, M. S. A., Webb, T. W. (2014). 'A Mechanistic Theory of Consciousness', *International Journal of Machine Consciousness*, 2. doi:10.1142/S1793843014400174.

Halligan, P. W., Fink, G. R., Marshall, J. C., Vallar, G. (2003). 'Spatial Cognition: Evidence from Visual Neglect,' *Trends in Cognitive Sciences*, 7, 125–33.

Horowitz, A. (2009). 'Attention to Attention in Domestic Dog (*Canis familiaris*) Dyadic Play, *Animal Cognition*, 12, 107–18.

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, 1220–6.

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 USA*, 103, 17048–52.

Kelly, Y. T., Webb, T.W., Meier, J. D., Arcaro, J., Graziano, M. S. A. (2014). 'Attributing Awareness to Oneself and to Others,' *Proceedings of the National Academy of Sciences USA*, 111, 5012–7.

Kentridge, R. W., Nijboer, T. C., Heywood, C. A. (2008). 'Attended but Unseen: Visual Attention is Not Sufficient for Visual Awareness, *Neuropsychologia*, 46, 864–9.

Koch, C., Tsuchiya, N. (2007). 'Attention and Consciousness: Two Distinct Brain Processes', *Trends in Cognitive Sciences*, 11, 16–22.

Lamme, V. A. (2004). 'Separate Neural Definitions of Visual Consciousness and Visual Attention: A Case for Phenomenal Awareness', *Neural Networks*, 17, 861–72.

Lau, H., Rosenthal, D. (2011). 'Empirical Support for Higher-Order Theories of Consciousness', *Trends in Cognitive Sciences*, 15, 365–73.

McCormick, P. A. (1997). 'Orienting Attention Without Awareness', *Journal of Experimental Psychology, Human Perception and Performance*, 23, 168–80.

Metzinger, T. (2010). *The Ego Tunnel*, New York: Basic Books.

Mysore, S. P., Knudsen, E. I. (2013). 'A Shared Inhibitory Circuit for Both Exogenous and Endogenous Control of Stimulus Selection', *Nature Neuroscience*, 16, 473–8.

Nisbett, R. E., Wilson, T. D. (1977). 'Telling More Than We Can Know – Verbal Reports on Mental Processes', *Psychological Review*, 84, 231–59.

Norman, L. J., Heywood, C. A., Kentridge, R. W. (2013). 'Object-Based Attention Without Awareness', *Psychological Science*, 24, 836–43.

Posner, M. I. (1980). 'Orienting of Attention', *Quarterly Journal of Experimental Psychology*, 32, 3–25.

Premack, D., Woodruff, G. (1978). 'Does the Chimpanzee Have a Theory of Mind?', *Behavioral and Brain Sciences*, 1, 515–26.

Saxe, R., Kanwisher, N. (2003). 'People Thinking About Thinking People: fMRI Investigations of Theory of Mind', *NeuroImage*, 19, 1835–42.

Scheidt, R. A., Conditt, M. A., Secco, E. L., Mussa-Ivaldi, F. A. (2005). 'Interaction of Visual and Proprioceptive Feedback During Adaptation of Human Reaching Movements', *Journal of Neurophysiology*, 93, 3200–13.

Skrbina, D. (2005). *Panpsychism in the West*, Cambridge: The MIT Press.

Szczepanowski, R., Pessoa, L. (2007). 'Fear Perception: Can Objective and Subjective Awareness Measures be Dissociated?', *Journal of Vision*, 10, 1–17.

Tononi, G. (2008). 'Consciousness as Integrated Information: A Provisional Manifesto', *Biological Bulletin*, 215, 216–42.

Tsushima, Y., Sasaki, Y., Watanabe, T. (2006). 'Greater Disruption Due to Failure of Inhibitory Control on an Ambiguous Distractor', *Science*, 314, 1786–8.

Vallar, G. (2001). 'Extrapersonal Visual Unilateral Spatial Neglect and its Neuroanatomy', *NeuroImage*, 14, 552–8.

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, 609–22.

van Swinderen, B. (2012). 'Competing Visual Flicker Reveals Attention-Like Rivalry in the Fly Brain', *Frontiers in Integrative Neuroscience*, 6, 96.

Webb, T. W., Graziano, M. S. A. (2015). 'The Attention Schema Theory: A Mechanistic Account of Subjective Awareness', *Frontiers in Psychology*. doi:10.3389/fpsyg.2015.00500.

Webb, T. W., Kean, H. H., Graziano, M. S. A. (2016). 'Effects of awareness on the control of attention'. *Journal of Cognitive Neuroscience*, 28, 842–51.

Wimmer, H., Perner, J. (1983). 'Beliefs About Beliefs: Representation and Constraining Function of Wrong Beliefs in Young Children's Understanding of Deception', *Cognition*, 13, 103–28.

Wolpert, D. M., Goodbody, S. J., Husain, M. (1998). 'Maintaining Internal Representations: The Role of the Human Superior Parietal Lobe', *Nature Neuroscience*, 1, 529–33.

Young, L., Dodell-Feder, D., Saxe, R. (2010). 'What Gets the Attention of the Temporo-Parietal Junction? An fMRI Investigation of Attention and Theory of Mind', *Neuropsychologia*, 48, 2658–64.