NOT FOR PUBLIC RELEASE

Michael S.A. Graziano

We Are Machines That Claim to Be Conscious

Abstract: *The attention schema theory explains how a biological, information processing machine can claim to have consciousness, and how, by introspection (by assessing its internal data), it cannot determine that it is a machine whose claims are based on computations. The theory directly addresses Chalmers' meta-problem of consciousness, the problem of why we think we have a difficult-to-explain consciousness in the first place.*

1. Introduction

Neuroscience has taught us that the brain is an information processing device. In the perspective that I take, and the theory I have suggested — the attention schema theory (AST) — we are information processing machines that, among other actions, make claims about ourselves (e.g. Graziano and Kastner, 2011; Graziano, 2013; Webb and Graziano, 2015). We claim to have something inside us, subjective experience, that is fundamentally non-physical. Logically, the brain cannot put out a claim unless it contains the information on which the claim is based. In my research, therefore, I have focused on the information set on which the claim of subjective experience is based. What cognitive purpose does it serve? What brain regions might be involved in constructing it? How is the machine engineered such that it makes that claim?

I would like to clarify at the outset what I mean by a non-physical property. If a person looks at a red apple, she not only processes information about the colour, but also claims to have a subjective experience of red — the 'what it feels like' component. One cannot push on subjective experience and measure a reaction force, scratch it

Correspondence:
Email: graziano@princeton.edu

Journal of Consciousness Studies, **26**, No. 9–10, 2019, pp. ?–?

and measure its hardness, or put it on a scale and measure its weight. It does not exist on those physical dimensions. In the sense of its physical non-measurability, subjective experience is non-physical, or even metaphysical in the strict sense of being above or outside the physical. This ethereal nature of subjective experience is precisely why it has been so difficult to understand.

But, objectively speaking, the phenomenon that faces us is much simpler. A brain-controlled agent constructs a self-description and on that basis makes claims about itself. There is no rational reason to suppose the claims are literally accurate. We already know from cognitive neuroscience that the brain constructs many internal models — bundles of information that represent items in the real world (Johnson-Laird, 1983; Holmes and Spence, 2004; Graziano, 2013). These models, whenever they have been studied in detail, are always simplified. They are quick-and-dirty descriptions, useful if not entirely accurate. The question in front of us is not: how does the brain generate a non-physical essence? Rather, we should ask: what set of information in the brain is the basis for our claim to have conscious experience, and what adaptive function does that information serve? AST does not explain how a brain generates a subjective experience. It explains how a machine makes claims about itself, and how the information on which those claims are based may have a cognitive, functional use.

Chalmers has written an insightful article, outlining what he has termed the hard problem and the meta-problem (Chalmers, 2018). One way to frame the hard problem is that consciousness is a private experience whose existence cannot be assessed from the outside. Because it cannot be physically measured, it cannot be scientifically studied. The meta-problem, in contrast, is the question of why we think we have a hard problem. Part of Chalmers' discussion focuses on an approach to consciousness called illusionism (Frankish, 2016). In that approach, consciousness does not exist as such — it is illusory. One of the earliest and most influential illusionist accounts is Dennett's idea of the user illusion (Dennett, 1992). Illusionism could be considered a proposed approach to the meta-problem — it suggests that we think we have a hard problem of consciousness because we are misinformed by an illusion.

AST specifically addresses Chalmers' meta-problem, because it addresses how a biological machine claims to have a hard problem. Yet in Chalmers' article, one senses his uneasiness over how to interpret AST. For example, he puzzles over the question: in AST, what

exactly is awareness? Is it an attention schema, or is it supposed to be an abstraction to which an attention schema refers? As I spell out in chapter 3 of my book *Consciousness and the Social Brain* (2013), AST does not easily pin down what, exactly, awareness itself is. The reason for the ambiguity, I believe, is that AST is fundamentally not a philosophical theory. It is an engineering theory. It explains the performance of a machine — it explains how a machine claims to have consciousness. It could be viewed as an illusionist theory, and is especially close to Dennett's account. Yet it may not perfectly fit into the illusionist category either — or at least it may provide a different emphasis. Illusionism seems to ask: how does the brain generate, if not an actual conscious experience, at least an illusory semblance of one? That framing focuses on how the brain generates something, and on consciousness as a distinct item of interest whose real or illusory nature can be debated. But in AST there is no meaningful answer to the question. Instead, the theory addresses how a machine makes claims, not how a machine generates experiences or illusions. We can understand how a car drives and a bird flies, from an engineering perspective. We should be able to understand, mechanistically, how a brain makes claims.

2. Model-Based Knowledge

To describe consciousness as the brain making claims, I acknowledge, sounds at first too reductive. But the crux of the argument lies in the information sets on which those claims are based. AST depends on model-based knowledge, as distinct from superficial knowledge. To explain what I mean, I will use an example that I have used in other recent accounts (Graziano, 2019).

Suppose a child plays at make-believe. She barks, crawls on all fours, and says, 'I'm a puppy!' Something in her brain contains the information that puppies bark and walk on all fours. Her brain has also constructed the proposition 'I'm a puppy!' or else she would not be able to make the claim. And yet that information exists in a larger context. Her brain contains a net of information including 'I'm not really a puppy', 'I'm making it up', 'I'm a little girl', and so on. Some of that information is present at a cognitive and linguistic level. Much of it is at a deeper, sensory or perceptual level. Her body schema is constructed automatically, beneath higher cognition, and describes the physical layout of a human body, not a puppy body. She sees her human hands in front of her, and the representations constructed in her

visual system confirm her human identity. She remembers eating breakfast with a spoon, going to school, reading a book — all human activities. The claim 'I'm a puppy' is superficial knowledge that is inconsistent with her deepest internal models.

But suppose I have the science fiction tools to manipulate the information in her brain. I alter her body schema to reflect the body of a puppy. I alter the information in her visual system and her memory to make it consistent with the puppy proposition. I remove the specific cognitive information that says 'I made that up to play a game'. I switch the information that says 'I'm certain this is not true' to its opposite. How would she know that she is not a puppy? Her brain is captive to the information it contains. Tautologically, it knows what it knows. She would no longer think of her puppy identity as a hypothetical. She would take it as a literal truth. There would be no reason for her to think otherwise. One might say that she now believes, intuitively, that she is a puppy; and here, to clarify the terminology, by 'believing something intuitively' I mean that her cognition is informed by deeper, automatically constructed, internal models. The belief, at the cognitive level, derives from the deeper internal models over which she has no cognitive control.

You could tell her, 'But you understand English. Puppies can't do that. Don't you think that suggests you've mistaken your identity?' If she is intellectually precocious, she might realize the logic of your argument. That new information, however, will be at a superficial, cognitive level. It will conflict with her deeper internal models. Like so many people, she will be in a position of believing one truth about herself intuitively, while entertaining a different truth intellectually.

Just so, I might be able to convince you intellectually that your claim to consciousness has its basis in an information set — an attention schema, as I'll explain in the next section. But intuitively, you still believe a different truth about yourself. When you rely on introspection — when your cognition accesses deeper internal models — they provide you with a different story. They inform you (incorrectly) that your consciousness is not just information or computation — it has a 'what it feels like' component, an ethereal essence dwelling inside you. Even if I have convinced you of my argument, you will find yourself conflicted, with superficial, intellectual knowledge pointing you towards one understanding and deeper, internal models, over which you have no cognitive control, anchoring you to a different understanding.

3. The Attention Schema

In this short piece, I will not give a complete account of AST or the supporting lines of evidence. I refer readers to previous publications (e.g. Graziano and Kastner, 2011; Graziano, 2013; Webb and Graziano, 2015). Instead, here, I briefly summarize the core concept.

Logically, we claim to have subjective experience for the same reason we make any claim — because the brain has constructed the requisite information on which the claim is based. Suppose a person looks at a red apple and reports having a subjective experience of red. It is not enough for the brain to construct colour information, which would allow the person to make the limited claim 'The apple is red'. We know, for example, that people who suffer from blindsight (Cowey, 2010) can process visual information and make claims about visual features, without reporting any conscious visual experience. To report a conscious experience, the brain must also construct the information on which it bases the claim 'I have something extra, a non-physical subjective experience, associated with the redness'.

The brain constructs descriptive sets of information because they act as useful models for real items in the world. One question facing us, therefore, is: what is the physically real item that is modelled by this particular information set, on which the claim of conscious experience is based?

For example, your brain constructs a set of information that is, moment by moment, correlated with the configuration of your right arm. That information set is called an arm schema, a part of the body schema (Graziano and Botvinick, 2002; Holmes and Spence, 2004; de Vignemont, 2018). It is the basis on which you can close your eyes and report on the presence and state of your arm. That internal model usually covaries with the physical arm, although the two can be dissociated. Like all internal models in the brain, the model of the arm is a detail-poor simplification, and can sometimes make errors and become misaligned. It *usually* describes the overall state of the arm. One could say that the fact that this particular set of signals in the brain co-varies with the state of the arm, by definition, makes it an arm schema. The close tracking of the arm is what makes it informative about the arm.

Can we find any physically real, objectively measurable item that co-varies with people's report of conscious experience? Yes. This question has a straightforward answer known in psychology and neuroscience for decades. The report of conscious experience tends to

co-vary with attention (e.g. Posner, 1994; Mack and Rock, 1998; Simons and Chabris, 1999; Cohen *et al.*, 2012). If a person reports being conscious of X, she is typically also attending to X. Attention and awareness can sometimes be separated. At least, attention without awareness has been demonstrated, though awareness without attention has not yet been convincingly shown (e.g. Kentridge, Heywood and Weiskrantz, 1999; Tsushima, Sasaki and Watanabe, 2006; Webb, Kean and Graziano, 2016). Separating the two depends on pushing the system to extremes, either through brain damage or laboratory conditions in which visual stimuli are degraded and presented at detection threshold. Most of the time, however, awareness closely tracks attention. (Indeed it seems to be easier to separate the arm from the arm schema than attention from awareness, at least in my experience having experimentally studied both topics.)

One might ask: is attention too narrow a phenomenon to cover subjective consciousness? Surely we are conscious of much more than we put at the focus of our attention. But objectively speaking, in decades of work on subjective awareness and attention, this intuition is not correct (e.g. Posner, 1994; Mack and Rock, 1998; Simons and Chabris, 1999; Cohen *et al.*, 2012). Awareness and attention co-vary most of the time. The confusion arises when people use a colloquial definition of attention, rather than a scientific one. In a typical colloquial definition, attention is a limited, central focus within the larger field of consciousness. In contrast, in neuroscience and psychology, attention is a process in the brain, primarily in the cerebral cortex, whereby a representation (such as a visual representation of an apple) has its signals enhanced, competing representations have their signals reduced, and the enhanced signals have a correspondingly greater impact on systems around the brain (Desimone and Duncan, 1995; Beck and Kastner, 2009). That enhancement can occur either due to greater external salience (bottom-up attention) or due to internal modulation (top-down control). Attention is not limited to one central object; it can be directed away from the fovea, for example, and it can be spread and divided. If you think that you are aware of something outside of your attention — that you are attending only to A while also aware of B, C, and D — that intuition is not correct; or at least, you are drawing on a colloquial definition of attention. By the scientific definition, you are probably attending to all of these items to some degree. Consciousness almost always co-varies with attention. It therefore effectively serves as a model of attention.

Chalmers suggests that linking consciousness to an attention schema is overly specific. Perhaps the brain constructs a general 'representation schema' which tells us what it means to represent information and gives rise to our claims about consciousness. But this suggestion stems from a misunderstanding of the theory. The report of conscious experience does not correlate with all representations in the brain. It correlates specifically with attention. Just so, the internal model of my arm is not a general 'moving object schema'. It is specifically an arm schema, because it tracks the state of my arm. Moreover, the functional use of an arm schema is to monitor, predict, and help control your arm; and the proposed functional use of an attention schema is to monitor, predict, and help control attention.

Suppose we were to design an attention schema from scratch. Our goal is to construct a useful information set descriptive of attention. For comparison, the arm schema contains stable information such as size, shape, jointed structure, and weight, as well as changing information such as how the arm is moving at the moment. Just so, the attention schema might describe both stable and changing properties of attention. Imagine a rich, textbook-style, scientific description of attention, including the details of the physical mechanisms present in the brain — and then imagine stripping from that description everything unnecessary for the brain to be informed about. We strip away information about neurons, synapses, inhibition and excitation — the physical truth of attention. We strip away information about bottom-up and top-down pathways, about fronto-parietal networks, about the thalamus and about the superior colliculus. We strip away information about the technical distinctions between exogenous and endogenous, engage and disengage, overt and covert, spatial and feature. We are left with a detail-poor description of attention as an amorphous 'thing' inside of me, a mental stuff that can grasp hold of objects in an abstract sense. The 'thing' can grasp hold of external objects like an apple, or internal objects like the thought that 2 + 2 = 4. The 'thing' has special powers such that, when it grasps hold of object X, it causes me to understand the details and the deeper meaning of X; it causes X to become vivid to me; it empowers me to choose to react to X, and to remember it for later. This stripped-down description of attention contains no information about the physical properties of the 'thing' inside me. As far as one can tell from the attention schema, that 'thing' lacks physicality.

My argument here is that if a brain uses the mechanism of attention, and if it constructs a simplified internal model of it, and if it makes

claims about itself on the basis of the information in that attention schema, then it ought to claim to have a subjective, non-physical, mental grasp, or experience, of objects. In this way, AST explains how a machine claims to have consciousness — without having to explain what consciousness itself is.

4. The Non-physical Essence

The philosopher François Kammerer asked an insightful question (Kammerer, 2016; 2018). Suppose AST is correct. The brain constructs an attention schema which represents general properties of attention, such as our ability to focus on and process information in depth. At the same time, it leaves out any depiction of the physical or mechanistic properties of attention. It does not specify that attention *lacks* a physical substance — it is merely silent on the topic. It is uninformative on the details of neurons and synapses. If our claims about consciousness derive from that internal model, then why do people typically make such a strong claim that consciousness is an ethereal essence, something inside of us that specifically *lacks* physical substance? Why do we not, instead, have an intuition of consciousness as an entity whose physical attributes — weight, size, hardness — are simply not yet known?

The answer may lie partly in a subtle distinction. I suggest that we do not generally understand consciousness as a thing whose physical dimensions are undetermined. Instead, we intuitively understand consciousness as something for which physical dimensions are *irrelevant*.

Imagine someone taps you on the shoulder. The touch activates skin receptors, and neuronal fibres transmit that information to the brain. Ultimately, your brain constructs a specific kind of internal model, a tactile model, a packet of information that describes that particular touch. The model contains information about the location of the touch, the intensity at onset, the pressure, the duration, the smooth or plush texture of a fingertip. It is a rich sensory representation. But it contains no information about taste. A touch on the shoulder does not come with a salty taste. I do not mean that a touch is bland and needs salt — no, it does not lie *anywhere* on any taste dimension. It does not occupy the same information space. Now that I have mentioned the possibility, you can consider it in a superficial, cognitive sense, but you cannot alter the deeper, internal model. Touch perception is an inborn process and is not open to cognitive modification.

If you could insert electrodes into a person's brain and read the information encoded in the tactile system, the perceptual model for a touch would presumably not contain the information 'And by the way, no taste is present'. It does not need the explicit negation. It is simply silent on taste. We do not intuitively understand touch to be something for which taste has been minimized; or something that might have a taste, but we just don't know yet what the taste is. Instead, we understand touch to be something for which taste is *irrelevant*.

I argue that the attention schema acts the same way. It depicts general properties of attention, but not physical, mechanistic properties. Based on that internal model, we intuitively believe in an inner mental experience that takes possession of information and drives action, the way attention does, but that has no specific relationship to physicality. Physicality is irrelevant to it. That mental essence is not physically graspable, smooth, textured, rough, bumpy, heavy, light, smelly, green, pointy — it does not lie anywhere on those physical dimensions, any more than a touch exists on the salty dimension.

And yet, in AST, the attention schema depicts at least one physical property. It depicts attention as having a physical location roughly inside us (see my prior accounts of the importance of localization in a model of attention: Graziano and Kastner, 2011; Graziano, 2013). Based on the information within that internal model, we should have an intuition about a mental essence that overlaps the physical world, in that you can point to a location and say 'it lives roughly here'. It is like a ghost, inhabiting physical space even as it lacks any relationship to other physical attributes. It has its own special power — to make us know and react. In this theory, the ghost in the machine, the consciousness inside us, is a topic of discussion among us only because our intuitions are informed by an attention schema, with its incomplete account of attention.

And so we come back to the hard problem and the meta-problem. In my proposed explanation, the belief in a hard problem derives from intuitions that come bubbling up from a deep, subsurface model, the attention schema. AST is a meta answer that explains why people believe in a hard problem in the first place.

References

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

Chalmers, D.J. (2018) The meta-problem of consciousness, *Journal of Consciousness Studies*, **25** (9–10), pp. 6–61.
Cohen, M.A., Cavanagh, P., Chun, M.M. & Nakayama, K. (2012) The attentional requirements of consciousness, *Trends in Cognitive Sciences*, **16**, pp. 411–417.
Cowey, A. (2010) The blindsight saga, *Experimental Brain Research*, **200**, pp. 3–24.
de Vignemont, F. (2018) *Mind and Body*, New York: Oxford University Press.
Dennett, D.C. (1992) *Consciousness Explained*, New York: Little-Brown.
Desimone, R. & Duncan, J. (1995) Neural mechanisms of selective visual attention, *Annual Review of Neuroscience*, **18**, pp. 193–222.
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.
Graziano, M.S.A. (2013) *Consciousness and the Social Brain*, Oxford: Oxford University Press.
Graziano, M.S.A. (2019) *Rethinking Consciousness*, New York: W.W. Norton.
Graziano, M.S.A. & Botvinick, M.M. (2002) How the brain represents the body: Insights from neurophysiology and psychology, in Prinz, W. & Hommel, B. (eds.) *Common Mechanisms in Perception and Action: Attention and Performance XIX*, pp. 136–157, Oxford: Oxford University Press.
Graziano, M.S.A. & Kastner, S. (2011) Human consciousness and its relationship to social neuroscience: A novel hypothesis, *Cognitive Neuroscience*, **2**, pp. 98–113.
Holmes, N.P. & Spence, C. (2004) The body schema and the multisensory representation(s) of peripersonal space, *Cognitive Processing*, **5**, pp. 94–105.
Johnson-Laird, P. (1983) *Mental Models*, New York: Lawrence Erlbaum.
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.
Kammerer, F. (2018) Can you believe it? Illusionism and the illusion meta-problem, *Philosophical Psychology*, **31**, pp. 44–67.
Kentridge, R.W., Heywood, C.A. & Weiskrantz, L. (1999) Attention without awareness in blindsight, *Proceedings of the Royal Society B: Biological Sciences*, **266**, pp. 1805–1811.
Mack, A. & Rock, I. (1998) *Inattentional Blindness*, Cambridge, MA: MIT Press.
Posner, M.I. (1994) Attention: The mechanisms of consciousness, *Proceedings of the National Academy of Sciences USA*, **91**, pp. 7398–7403.
Simons, D.J. & Chabris, C.F. (1999) Gorillas in our midst: Sustained inattentional blindness for dynamic events, *Perception*, **28**, pp. 1059–1074.
Tsushima, Y., Sasaki, Y. & Watanabe, T. (2006) Greater disruption due to failure of inhibitory control on an ambiguous distractor, *Science*, **314**, pp. 1786–1788.
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., Kean, H.H. & Graziano, MS.A. (2016) Effects of awareness on the control of attention, *Journal of Cognitive Neuroscience*, **28**, pp. 842–851.