Graziano, MSA (2020) What makes us so certain that we're conscious? Cognitive Neuroscience, 12: 67-68, https://doi.org/10.1080/17588928.2020.1838468.

COGNITIVE NEUROSCIENCE
https://doi.org/10.1080/17588928.2020.1838468

Routledge
Taylor & Francis Group

What makes us so certain that we're conscious?

Michael S. A. Graziano

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

ABSTRACT

In the attention schema theory (AST), having an automatically constructed self-model that depicts you as containing consciousness makes you intuitively believe that you have consciousness. The reason why such a self-model evolved in the brains of complex animals is that it serves the useful role of modeling, and thus helping to control, the powerful and subtle process of attention, by which the brain seizes on and deeply processes information.

ARTICLE HISTORY
Received 27 August 2020
Revised xx xxx xxxx
Published online xx xxx xxxx

KEYWORDS
Attention; awareness; consciousness; self-model

The field of consciousness studies is crowded. The theories multiply and the data do not yet distinguish clearly among them. I recently argued that many of the current theories are related to each other; they are partial glimpses of deeper ideas, and not necessarily all rivals (Graziano et al., 2019). Yet we are still left with the task of sorting and evaluating them. Doerig et al. (2020) have now proposed an excellent, preliminary set of logical criteria for evaluating this profusion of theories. The systematic approach is needed and much appreciated.

The theory that my colleagues and I proposed, the attention schema theory (AST), is one of those evaluated by Doerig et al., and we appreciate the clear and accurate summary of AST and the insightful discussion. Here I will focus microscopically on one aspect of AST that is often misunderstood. Doerig et al. mention that aspect, but do not go into enough detail either to get it wrong or right. It is a seemingly minor point. And yet, in the end, it may be the most important part of AST.

In AST, the brain constructs a model of its own attention, to help monitor, predict, and thus control attention, much as the brain constructs a body schema to help monitor, predict, and thus control the physical body (Webb & Graziano, 2015). We know we have a physical body – we know it in an immediate, intuitive way – because of the information contained within the body schema (Graziano & Botvinick, 2002). Without a body schema, we would have only an intellectual, abstract knowledge of our physical selves. Just so, in AST, we know we have a mysterious power inside us that mentally possesses items with vividness and immediacy, and enables us to decide and react to those items – we know it in a direct, intuitive way – because of the information contained within an attention schema. That information, by depicting the process of attention without depicting any of the mechanistic details, both informs us and, in a sense, misinforms us. The model of attention, being more efficient than accurate, tells us that we have a kind of magical or nonphysical power inside us.

At the heart of the theory is something I consider to be a logical certainty: everything we claim about ourselves, everything we think we know, no matter how certain we are, depends on information in the brain that tells us that it is so, or we wouldn't be able to have the thought or make the claim. And the brain's models of itself, and of the world, are never fully accurate. They are only models, and are partial, cut-corner, and sometimes superficial, for efficiency. The consciousness we intuitively 'know' we have is – almost certainly, I would argue – a cartoonish distortion of a property we actually have. And the property we actually have, from which our claims of consciousness derive, I suggest, is mechanistic attention.

Doerig et al. ask of AST, 'If there is no computational explanation of what really counts as attention modeling, the other systems problem arises: what is crucial for consciousness in the human implementation of attention modelling? Which other systems have it and are thus predicted to be conscious?'

Others have asked a similar question. To paraphrase, 'I can easily build a neural network that uses attention and that models its own attention. It would require only a few lines of code. Is that toy network conscious?'

Not according to AST. An attention schema is not a magical amulet that produces consciousness.

Suppose a machine has a model of attention; the model is limited, and contains only information about where attention is currently pointed. Such a model may

CONTACT Michael S. A. Graziano graziano@princeton.edu Department of Psychology and Neuroscience, Princeton University, Princeton, NJ 08544

be useful for the control of attention. But it will not inform the machine that it is conscious. Imagine we have a perfect speech engine – a black box that takes in all types of information and translates it into English. If we plug that speech engine into the machine's attention schema, what will come out? 'Over here! Over there! X, Y and Z equals this, that, and the other!' The machine will say nothing about consciousness – the property is irrelevant to it.

Suppose the machine has a much richer model of attention. Somehow, attention is depicted by the model as a Moray eel darting around the world. Maybe the machine already had need for a depiction of Moray eels, and it coapted that model for monitoring its own attention. Now we plug in the speech engine. Does the machine claim to have consciousness? No. It claims to have an external Moray eel.

Suppose the machine has *no* attention, and no attention schema either. But it does have a self-model, and the self-model richly depicts a subtle, powerful, nonphysical essence, with all the properties we humans attribute to consciousness. Now we plug in the speech engine. Does the machine claim to have consciousness? Yes. The machine knows only what it knows. It is constrained by its own internal information.

AST does not posit that having an attention schema makes one conscious. Instead, first, having an automatic self-model that depicts you as containing consciousness makes you intuitively believe that you have consciousness. Second, the reason why such a self-model evolved in the brains of complex animals, is that it serves the useful role of modeling attention. The first part I consider to be a logical certainty and the essential solution to the problem of consciousness. The second part is a specific hypothesis, that, in my view, is increasingly well supported by evidence, especially the evidence on the relationship between attention and awareness (Wilterson et al., 2020).

Disclosure statement

No potential conflict of interest was reported by the author.

References

Doerig, A., Schurger, A., & Herzog, M. H. (2020). Hard cfriteria for empirical theories of consciousness. *Cognitive Neuroscience*. https://doi.org/10.1080/17588928.2020.1772214

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

Graziano, M. S. A., Guterstam, A., Bio, B. J., & Wilterson, A. I. (2019). Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. *Cognitive Neuropsychology*. https://doi.org/10.1080/02643294.2019.1670630

Webb, T. W., & Graziano, M. S. A. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in Psychology*, *6*, article 500. https://doi.org/10.3389/fpsyg.2015.00500

Wilterson, A. I., Memper, C. M., Kim, N., Webb, T. W., Reblando, A. M. W., & Graziano, M. S. A. (2020). Attention control and the attention schema theory of consciousness. *Progress in Neurobiology*, *101844*. https://doi.org/10.1016/j.pneurobio.2020.101844