Consciousness is already solved: The continued debate is not about science.

(Comment on a target article by Bjorn Merker, Kenneth Williford, and David Rudrauf, The Integrated Information Theory of consciousness: A case of mistaken identity. Behavioral Brain Sciences, in press, 2022.)

Michael S. A. Graziano

Princeton University

Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08544

Tel: 609 258 7555

Email: Graziano@princeton.edu

Web page: https://grazianolab.princeton.edu/

Abstract

A logical explanation of consciousness has been known for decades. The brain must construct a specific set of information about conscious feeling (theory-of-mind information), causing people to believe, think, and claim to have consciousness. Theories that propose an actual, intangible feeling are non-explanatory. They add a magical red herring while leaving unexplained the objective phenomena: the believing, thinking, and claiming.

Main Text

The integrated information theory (IIT), a popular theory of consciousness, has conceptual flaws, and the article by Merker *et al.* (2021) brilliantly dismantles it. I wholeheartedly agree with the critique. Moreover, I admit to some frustration with the field of consciousness studies. The explanation of consciousness has been known for decades. It is as though, on the one hand, we have the logic that 2+2=4, while on the other hand, a public debate still rages. The 2+2 argument was described at least as far back as Dennett (1991) or even Nisbett and Wilson (1977). It has been called illusionism (Frankish, 2016), though I argue the name is misleading (Graziano, 2019). To explain the 2+2 argument, I will outline two principles.

Principle 1: Information that comes out of a brain must have been in that brain. Perhaps we can call it a computational conservation of information. Logically, nobody can think, believe, or insist on any proposition unless that proposition is represented by specific information in the brain – and in a form that affects the systems responsible for thinking, believing, and claiming.

If you believe that you have a subjective, phenomenal experience – an experience of some of the information content in your head – then that belief obeys principle 1. You think it, believe it, and claim it, because your brain contains information descriptive of it.

"But what causes the experience, the *feeling*?"

You have information that tells you that you have a feeling. That is why you believe you have a feeling.

"No, I definitely have the feeling. I know it, because I can feel it right now."

That argument is tautological. To argue for the presence of feeling because you feel it, is to state, "it's true because it's true." To query whether you have a subjective feeling, your cognition accesses data, and the data constrains your belief and your answer. The presence of *information about feeling* fully explains the phenomenon.

By principle 1, the question of consciousness becomes: why does the human brain (and perhaps the brains of other animals) construct a specific set of information that describes an intangible property of phenomenal experience? Put this way, the question is lifted out of the domain of unsolvable mystery and into biological and evolutionary significance. The brain presumably constructs information sets for specific, adaptive reasons. Why does it construct this particular one?

Principle 2: The brain builds information sets, or models, of reality; but the models are never fully accurate or detailed. In every case in which the comparison has been made, the model differs from the item being modeled. The paradigmatic example is color, for which the visual system constructs a greatly simplified model of the complex reflectance spectra of surfaces.

Suppose that (on principle 1) your brain constructs a model, a root set of information from which it derives the claim that you have a conscious feeling. Now suppose that the model is not an empty illusion, but represents something physically real inside you. What is that physically real thing? By principle 2, the model differs from the real item. Science is under no obligation to look for, or to explain, a physical process that has exactly the same properties as the phenomenal "feeling" that you think and believe you have. In analogy, when the police draw a sketch of a suspect, nobody looks for a man whose face is made of gray pencil lines. We look for the man *represented by* the simplified caricature. Just so, the scientific task is to look for a physical brain process for which your introspective claims of conscious experience act as a crude, simplified model.

In my work, I hypothesized that the physically real item in question is attention. Selective attention is when a subset of information, especially in the cerebral cortex, is processed deeply and at high signal strength, thereby impacting downstream systems. In the theory, you believe you have a subjective experience of something; the belief derives from an automatically constructed model; the model is a detail-poor, schematic representation of your attention on that thing; and modeling one's own attention is necessary to control attention. This theory is called the Attention Schema Theory (AST). I have argued for it based on data on the relationship between attention and awareness, neuroscientific data, and, most recently, artificial neural networks (Graziano and Kastner, 2011; Kelly *et al.*, 2014; Webb and Graziano, 2015; Webb *et al.*, 2016; Wilterson *et al.*, 2020; Wilterson and Graziano, 2021; Wilterson *et al.*, 2021).

One does not need to accept AST in specific, however, to realize that consciousness has already been explained, in general, by principles 1 and 2. Only by ignoring logic can any other class of explanation remain standing. IIT, like so many theories of consciousness, violates both principles as follows.

IIT purports to explain the presence of a conscious feeling. In contrast, by principle 1, science can explain how a specific set of information about conscious feeling (theory-of-mind information) is constructed in the brain, causing people to believe, think, and claim to have consciousness. If you posit the presence of an actual feeling, whatever that additional essence might be, it is non-explanatory. It leaves unexplained the objectively known phenomena: the believing, thinking, and claiming.

In IIT, to discover the source of consciousness, one must start with the specific properties that people introspectively believe to be present in consciousness – such as unity and richness – and find something measurable that has those same properties. In contrast, in principle 2, the introspected information set is not literally correct, but is instead a crude, schematic model of something else. Science is not obliged to find a real entity that perfectly matches what the schematic model tells us we have. Rather, our scientific job is to identify the real, physical, brain process that is imperfectly modeled by information in the brain, such that, on the basis of that crude information, we believe we have consciousness.

Conflict of interest statement: The author has no conflict of interest with respect to this article.

Funding statement: This work was supported by the Princeton Neuroscience Institute Innovation Fund.

References

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

Frankish, K. (2016). Illusionism as a theory of consciousness. *Journal of Consciousness Studies, 23,* 1-39.

Graziano MSA, Kastner S (2011). Awareness as a perceptual model of attention. Cognitive Neuroscience, 2, 125-133.

Kelly YT, Webb TW, Meier JD, Arcaro MJ, Graziano MSA (2014). Attributing awareness to oneself and to others. Proceedings of the National Academy of Sciences USA, 111, 5012-5017.

Merker B, Williford K, Rudrauf D (2021). The Integrated Information Theory of consciousness: A case of mistaken identity. Behavioral and Brain Sciences, In press.

Nisbett RE, Wilson TD (1977). Telling More Than We Can Know - Verbal Reports on Mental Processes. Psychological Review, 84, 231-259.

Webb TW, Graziano MSA (2015). The attention schema theory: a mechanistic account of subjective awareness. Frontiers in Psychology, 6, article 500, doi: 10.3389/fpsyg.2015.00500.

Webb TW, Kean HH, and Graziano MSA (2016). Effects of awareness on the control of attention. Journal of Cognitive Neuroscience, 28, 842-851.

Wilterson AI, Graziano MSA (2021). The Attention Schema Theory in a Neural Network Agent: Controlling Visuospatial Attention Using a Descriptive Model of Attention. Proceedings of the National Academy of Sciences USA, In press.

Wilterson AI, Kemper CM, Kim N, Webb TW, Reblando AMW, Graziano MSA (2020). Attention control and the attention schema theory of consciousness. Progress in Neurobiology, 195, doi: 10.1016/j.pneurobio.2020.101844.

Wilterson AI, Nastase SA, Bio BJ, Guterstam A, Graziano MSA (2021). Attention, awareness, and the right temporoparietal junction. Proceedings of the National Academy of Sciences USA, 118, e2026099118, doi: 10.1073/pnas.2026099118. PMID: 34161276.