Published in *Philosophical Psychology* (2023) 36: 438–441
doi:10.1080/09515089.2022.2092465
mark.sprevak@ed.ac.uk

Understanding phenomenal consciousness while keeping it real

Mark Sprevak
University of Edinburgh

19 June 2022

ANIL SETH, *Being You*. London: Faber & Faber, 2021, 378 pp., £20.00 (hardback). ISBN 9780571337705.

Being You is a beautifully written book about consciousness. It pulls off the seemingly impossible trick of being a deeply personal account of the author's own life and conscious experiences while also being a substantial contribution to the philosophy and science of consciousness. On the latter score, the intentions of this book are broad and programmatic. It sketches an agenda for a pragmatic, scientific approach to consciousness that has a different character to more familiar research programmes (e.g. the global workspace theory or proposals about higher-order thought). The book is interspersed with reports from empirical research studies and well-chosen examples from literature, film, and the arts. Seth is exceptionally good at explaining complex scientific ideas in clear, concise, and accessible ways. Particular highlights are his introductions to integrated information theory and the free-energy principle – I would recommend these as the first things someone new to these topics should read.

Seth says that his book is about what he calls the 'real problem' of consciousness, which he distinguishes from the hard problem of consciousness. The hard problem is to explain how and why phenomenal consciousness arises inside a physical world. The real problem takes the brute existence of phenomenal consciousness for granted – it does not attempt to explain how our experience arises from purely physical ingredients. Instead, it suggests that a science of consciousness should aim to give us the tools to *explain*, *predict*, and *control* our phenomenal experiences. This may

include searching for neural correlates of phenomenal consciousness as well as building computational theories that capture general principles about how physical goings on in our brains and bodies are systematically linked with certain types of conscious feeling. Seth argues that science sometimes offers us an understanding of a mysterious target phenomena, not by reducing it to something else, but by making it more predictable, weaving it into our wider explanatory practices, and placing it under our instrumental control. These broader scientific goals – which may or may not be accompanied by a successful metaphysical reduction – are often obscured in philosophy by an overly narrow focus on the hard problem.

Focusing on the real problem of consciousness is not, however, intended to mean that we give up on the hard problem, or at least, not forever. Seth suggests that tackling the real problem could provide an indirect means of addressing the hard problem. The idea is that, as we acquire more powerful models which allow us to explain, predict, and control phenomenal consciousness, phenomenal consciousness will, by degrees, become less puzzling to us. In the limit – if phenomenal consciousness and its relationship to the physical world were *completely* modelled by a quantitative, well-confirmed scientific theory – its brute occurrence might begin to seem completely ordinary and unexceptional to us, at which point 'the hard problem will fade away, disappearing in a puff of metaphysical smoke' (p. 28). Seth claims that something similar happened to the phenomenon of *being alive*. The question of which physical/functional properties are sufficient for *being alive* no longer strikes us as a particularly important or urgent problem for science. This is not because we have successfully reduced *being alive* to a set of known physical/functional conditions (we have not), but because we now have an extensive grasp of how to explain, predict, and control many of its associated phenomena. Solving the real problem may undercut the motivation for an ontological reduction.

Seth suggests that the real problem of consciousness should be divided into three separate questions concerning how to explain, predict, and control the *level* of conscious experience, the *contents* of phenomenal consciousness, and our feelings of *self* consciousness. I will sketch his suggestions for each.

The idea that phenomenal consciousness comes in levels is a common starting point for many aspects of work on consciousness. There appears to be an obvious difference between the amount of phenomenal consciousness inside you under general anaesthesia versus when you are fully awake. There also appears to a difference between the amount of phenomenal consciousness inside a rock versus that in a lively and inquisitive five-year-old child. Seth suggests that one goal for a science of consciousness is to make these informal ideas more precise: to develop practical and operational numerical measures of the presence of phenomenal consciousness, identify the neural and functional features associated with them, and bring all this

under our control via interventions such as anesthesia and pharmaceuticals.

Seth describes three candidate measures for the level of phenomenal consciousness. These are PCI (fire a magnetic pulse into the brain, measure how compressible are data about its electrical ripples); his own 'causal density' measure (based around the econometric concept of Granger causality); and integrated information theory's Φ. Regarding the last measure, Seth distances himself from some of the bold metaphysical claims of integrated information theory (IIT). IIT is normally taken as saying that Φ is necessarily connected with the occurrence of phenomenological consciousness. Seth suggests that Φ, like PCI and causal density, may only be contingently connected to consciousness, and hence it may not be a perfectly accurate measure of the level of consciousness. In principle, a system might have high Φ and no conscious experience, or conscious experience and low Φ. This allows Seth to avoid some of the counterintuitive consequences of IIT (e.g. that an inactive set of logic gates could be conscious), while still allowing that Φ plays a useful role in science alongside other measures such as PCI and causal density. He leaves it open whether there is one single 'true' measure of the level of phenomenal consciousness (as advocates of IIT claim for Φ). In line with the pragmatic character of his project, Seth suggests that this isn't something that can be judged ahead of enquiry.

Regarding the contents of consciousness, the book's key idea is that perception is 'controlled hallucination'. According to Seth, this is shorthand for a phenomenologically souped-up version of predictive processing. Predictive processing (aka 'predictive coding') is a theory about the subpersonal computational processes by which our brains achieve perception, cognition, and action. Seth's 'controlled hallucination' view assumes that the predictive processing story is basically correct but it adds an additional claim: certain elements of the subpersonal predictive processing architecture *are* the contents of our conscious experience. Specifically, Seth identifies the contents of consciousness with the contents of our best, stable top-down predictions:

> What we perceive is given by the content of all the top-down predictions together, once sensory prediction errors have been minimised … as far as possible. (p. 106)
>
> … top-down predictions do not merely bias our perception. They *are* what we perceive. Our perceptual world alive with colours, shapes, and sounds is nothing more and nothing less than our brain's best guess of the hidden causes of its colourless, shapeless, and soundless sensory inputs. (p. 115)

It is worth noting that this identity claim isn't fully justified inside the book. Philosophers may be surprised by this: they tend to expect certainty and watertight

arguments to defend the claim that $A = B$. Seth's proposal should be understood more in the pragmatic spirit of a conjecture or working hypothesis to guide and structure scientific enquiry regarding the real problem. His idea is that we should avail ourselves of a powerful quantitative framework – predictive processing – to try to better explain, predict, and control our phenomenal consciousness. There may be other approaches, and this approach may ultimately prove to be unsuccessful, but its justification lies not in evidence marshalled today, but in its eventual success or failure.

That said, it is worth flagging a few doubts about the proposal. Here are three that I had. First, according to predictive processing, good stable top-down predictions appear at many different levels in the neural hierarchy from retinal ganglion cells to the highest anatomical reaches of the cortex. However, only a tiny minority of top-down predictions appear to show up in the contents of our phenomenal consciousness. What makes the contents of some top-down predictions appear in phenomenal consciousness and others not? Second, the top-down predictions are assumed to be probabilistic, but our phenomenal experience appears to be categorical (the world seems a particular way to us). If the contents of the two are the same, what explains this mismatch? Third, our subjective experiences of surprise, alarm, and confusion appear to have a distinctive phenomenology, yet they seem to implicate bottom-up error signals rather than top-down predictions. Why then don't the contents of error signals also contribute to determining the contents of phenomenal consciousness? One might ask a similar question about precision weighting. Precision weighting scales error signals up and down and it is associated with attentional focus and with uncertainty about the reliability of sensory signals. Uncertainty and attentional shifts often have particular phenomenological feelings associated with them. Why not allow that these other aspects of the prediction-error-minimisation machinery – distinct from top-down predictions – may contribute to the contents of phenomenal consciousness?

Seth applies his theory to affective experiences, such as emotions and moods. He argues that affective states are top-down predictions about interoceptive sensory data – for example, about our blood pressure, temperature, heartbeat, breathing, and gastric tension. Those physiological variables do not show up directly in the contents of our conscious experience (just as the frequency and amplitude of light waves hitting our retinas do not show up directly in consciousness). Instead, they cause signals to travel up from our internals organs towards the brainstem and thalamus, and these signals are met, and attempted to be cancelled, by top-down predictions originating in the cortex. It is these top-down predictions from the cortex that determine the contents of consciousness. Our affective experiences are top-down predictions about interoceptive data. The conscious state of *being afraid* is a top-down prediction from our cortex that best explains the signals that arrive

at the brainstem arising from an elevated heart rate, breathing, etc. (and that best fits with our priors and other predictions based on, e.g., visual evidence). In Seth's language, and taking a cue from Descartes, we are 'beast machines': organisms that are tuned to predict (and regulate) our body's physiological variables. Our brains are not organs of pure rationality but predictive regulators that are tuned to the particularities of our bestial bodies.

In the final part of the book, Seth applies this 'beast machine' idea to our subjective experience of ourselves as a persisting agent over time. Despite the many changes that our bodies and minds undergo over a lifetime, we still have a subjective experience of being the same person – of 'being you'. What explains this? Seth suggests that it is another top-down prediction, albeit a higher-level one, that reflects a pattern of stability in our bestial, interoceptive predictions. In order for us to exist, our key physiological variables have to remain within certain ranges – if our heartbeat were to stop, or our breathing were to cease, we would die. Hence, given that we exist, there has to be an inherent stability to our bestial predictions. This stability is compounded by the action-determining character of those predictions. If a predictive system decides to hold onto a top-down prediction with a high degree of confidence (if it 'clamps' it as true), then other variables in the system will tend to change in preference, e.g. it might revise other predictions or cause the environment to change, so as to keep the clamped prediction true. In this way, a prediction that is accorded a high degree of certainty may serve as a kind of self-fulfilling expectation. Seth suggests that we predict that we are the same over time, not only because our key physiological variables tend to remain within the same range over time, but also because by having a strongly held prediction that they do remain the same, we help to keep them within that range. Your subjective sense that there is some 'you' that is stable and persisting over time arises because of the doubly-stable nature of your top-down interoceptive predictions.

Seth ends his book on a cautionary note. Even if his ideas concerning the real problem prove correct, important gaps would still remain in our understanding of consciousness. He considers what would happen if the predictive processing algorithm were run inside an electronic computer. He conjectures that it would be unlikely to be accompanied by phenomenal experience. He tentatively concludes that there is something special about our material biological make-up that, combined with the prediction generation machinery, breathes consciousness into us. What that special ingredient is, however, is left open and unexplored. In the end, the hard problem – the search to understand why certain physical conditions give rise to phenomenal consciousness – is a difficult one to escape.