Predictive processing as an empirical theory for consciousness science

Article (Accepted Version)

Predictive processing as an empirical theory *for* consciousness science

Anil K Seth[1,2] **and Jakob Hohwy**[3]

[1] Sackler Centre for Consciousness Science and School of Engineering and Informatics, University of Sussex, BN1 9QJ, UK
[2] CIFAR Program on Brain, Mind, and Consciousness, Toronto, ON, Canada
[3] Cognition & Philosophy Lab, Monash University, Melbourne, Australia

Abstract

The theories of consciousness discussed by Doerig and colleagues tend to monolithically identify consciousness with some other phenomenon, process, or mechanism. But by treating consciousness as singular explanatory target, such theories will struggle to account for the diverse properties that conscious experiences exhibit. We propose that progress in consciousness science will best be achieved by elaborating systematic mappings between physical and biological mechanisms, and the functional and (crucially) phenomenological properties of consciousness. This means we need theories *for* consciousness science, perhaps more so than theories *of* consciousness. From this perspective, 'predictive processing' emerges as a highly promising candidate.

Main text

An admirable driver for Doerig et al.'s approach is to let empirical considerations determine the assessment of ToCs, however, as philosophers know well, shedding conceptual and metaphysical presuppositions is difficult (e.g., some kind of conceptual "definition" is needed to make any "observations", and, letting input-output profiles determine the fate of ToCs implicitly favours some kind of functionalism). Whereas there is thus plenty to say about how philosophically neutral their four main criteria are, the strength in Doerig et al.'s approach lies in collating them and jointly testing ToCs against them.

How do ToCs measure up, then? Doerig et al. are pessimistic, blaming the ToC "quagmire" on each theory's propensity for identifying one type of phenomenon, process or mechanism with consciousness. Their pessimism seems at odds with Table 1, which appears positively pollyannaish: most ToCs only fail on one criterion (often 'unfolding', which has been compellingly criticised (Negro, 2020)), or 'small network', which seems answerable by fully developed theories), and they deliver diverging answers (or are yet to develop answers) to questions about phenomenological "observations" that remain under active discussion (e.g., unity of consciousness).

Nevertheless, we agree with Doerig et al. that many ToCs err in monolithically associating consciousness with just one favoured process or mechanism. We believe this is the case because being monolithic leaves theories poorly placed to (contrastively) explain conscious phenomenology (as opposed to its mere presence or absence). An alternative perspective, which it seems Doerig et al. would be sympathetic to, is that a useful theory need not identify consciousness with something else, but may instead provide systematic mappings

between properties of consciousness and underlying mechanisms, where the relevant properties of consciousness can be both functional and (critically) phenomenological. This perspective is superficially less ambitious, because it does not require that the underlying mechanisms be either necessary or sufficient for consciousness. But it is more tractable, since it replaces the single grand mystery of 'solving consciousness' with a series of smaller challenges concerned with explaining why conscious experiences are the particular way they are (e.g., why a visual experience of an object feels different from an emotional experience). One can think of this strategy as elaborating the approach of searching for 'neural correlates of consciousness' (Koch, Massimini, Boly, & Tononi, 2016) where candidate correlates should now also be able to explain, predict, and control the phenomenological properties with which they correlate (Hohwy & Seth, (submitted); Seth, 2009).

What other candidate theories might there be? Here, we briefly advertise a different perspective on a theory only mentioned in passing by Doerig et al.: predictive processing [PP, (Clark, 2013; Hohwy, 2013)]. PP views perception as a process of inference, so that perceptual content is constituted by the brain's 'best guess' of the (hidden) causes of sensory input. Elaborations of vanilla PP into action (active inference) and interoception (interoceptive inference) highlight how conscious contents are shaped by action, and how experiences of emotion and selfhood may be underpinned by neural predictions geared more towards control and regulation than towards discovering the world (or body) 'as it is' (Seth & Tsakiris, 2018). In these and other ways, PP furnishes a rich and empirically-oriented set of resources for establishing systematic mappings between biological mechanisms and the functional and phenomenological properties of consciousness (Hohwy & Seth, (submitted)), not all of which need be replicated in all sizes and types of systems (e.g., allostasis, self-evidencing, deep temporal counterfactualising or explicit self-modelling). In this sense, PP is not a theory of consciousness: it is a theory *for* consciousness science.

Doerig et al. briefly mention very recent attempts to use PP as an explicit theory of consciousness (as 'PPT'). However, had it been more extensively discussed as a theory *for* consciousness, it would have sat uneasily with their "paradigm case" criterion, which worries that theories that do not directly address the target problem of consciousness may instead be addressing co-occuring processes, rather than consciousness *per se*. Indeed, PP does not directly address consciousness, but we see this as a potential strength rather than a weakness. The key point is that PP can be strongly linked to consciousness by adopting phenomenological (and functional) properties of consciousness as explanatory targets. If a mechanism phrased in the language of PP generates testable predictions about such properties, it is doing useful work for consciousness science.

Finally, one criterion that Doerig et al imply but do not explicitly highlight is that a candidate ToC should be experimentally testable. This criterion is implicit in the title of their target article, but it deserves amplification in at least two ways. First, current ToCs vary dramatically in their testability. Some theories – such as global workspace theory – make predictions that are readily testable with current methods. Others – such as integrated information theory (Tononi, Boly, Massimini, & Koch, 2016) – are more difficult to test, certainly in their fine details. Second, testability is not the same as falsifiability. It may be too much to ask of a ToC (or any theory) that it is falsifiable in a Popperian sense. Rather, a useful ToC ought to be experimentally fecund in the sense that it gives rise to a steady

stream of testable predictions that collectively build explanatory insight. This perspective on the philosophy of science – due primarily to Lakatos (Lakatos, 1978) – is also well satisfied by taking PP as a theory *for* consciousness science, rather than as a theory *of* consciousness.

References

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav Brain Sci, 36*(3), 181-204. doi:10.1017/S0140525X12000477

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hohwy, J., & Seth, A. K. ((submitted)). Predictive processing as a systematic basis for identifying the neural correlates of consciousness.

Koch, C., Massimini, M., Boly, M., & Tononi, G. (2016). Neural correlates of consciousness: progress and problems. *Nat Rev Neurosci, 17*(5), 307-321. doi:10.1038/nrn.2016.22

Lakatos, I. (1978). *The methodology of scientific research programmes: Philosophical papers*. Cambridge: Cambridge University Press.

Negro, N. (2020). Phenomenology-first versus third-person approaches in the science of consciousness: the case of the integrated information theory and the unfolding argument. *Phenomenology and the Cognitive Sciences*. doi:10.1007/s11097-020-09681-3

Seth, A. K. (2009). Explanatory correlates of consciousness: Theoretical and computational challenges. *Cognitive Computation, 1*(1), 50-63.

Seth, A. K., & Tsakiris, M. (2018). Being a Beast Machine: The Somatic Basis of Selfhood. *Trends Cogn Sci, 22*(11), 969-981. doi:10.1016/j.tics.2018.08.008

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). Integrated information theory: from consciousness to its physical substrate. *Nat Rev Neurosci, 17*(7), 450-461. doi:10.1038/nrn.2016.44