ASSESSING CRITERIA FOR THEORIES

David Rosenthal
Philosophy and Cognitive Science
Graduate Center, City University of New York

Penultimate Version

Abstract: I raise concerns about Doerig et al's general project, about three of their criteria, and about their treatment of higher-order-thought theory.

I. Doerig et al (forthcoming; henceforth DSH) believe that scientific findings about consciousness are now rich enough to pare down the many current theories of consciousness (ToCs), and that their "hard criteria" will force ToCs to confront those empirical findings. But the proliferation of ToCs is likely due less to a failure to address empirical findings than to a lack of clarity about exactly what phenomenon the ToCs are examining.

DSH sidestep that issue, writing that "consciousness seems to evade a rigid definition, making it difficult to pit theories against each other like in other scientific disciplines." Still, ToCs must delineate the phenomenon they mean to explain. Perhaps because DSH see such delineation as involving a "rigid definition," they caricature it as something "philosophical" (sometimes "metaphysical"!), which they can safely ignore. But having a usable delineation of the target phenomenon is not philosophy; it's just good science. Without that no theory is evaluable, and theories can just talk past one another.

DSH urge that the contrast between being awake and asleep and experimental work on masking provide "clear empirical phenomena," which obviate the "need to posit a theoretical definition" (1). But they don't obviate the need to delineate the phenomenon under examination. The many differences between being awake and asleep and between masked and unmasked perceiving are far too varied to help with that.

Those two contrasts, moreover, involve distinct phenomena. Being awake consists in a creature or system's being conscious. Masking, in contrast, involves a perceptual content's not being conscious. And many unconscious contents occur while an individual is conscious. DSH do note that the consciousness of systems and of contents (§II.1) are different, but still often elide the two.

DSH insist that "as for any scientific theory, the proposed mechanism [for a ToC] must be both necessary and sufficient to explain data about consciousness" (4). That's puzzling. Scientific theories do sometimes propose necessary and sufficient conditions, but often just one or the other. It's unclear what motivates this demand.

DSH caution that their criteria don't address "specific ToCs," but only "common ideas and principles of ToCs" (2; their emphasis). But details can make all the difference; one

can't accurately assess a theoretical approach using just generalities. DSH note that "[c]urrently, there are very few mutual comparisons between ToCs" (2). Still, some authors do address competing theories. DSH might have amplified on that, taking due account of relevant details.

II. DSH's first criterion is very important; a ToC must explain the contrast between consciousness and its absence. But the other three criteria are problematic, in part because their abstract character blocks the promised connections with specific empirical findings.

The second criterion appeals to the "mathematical fact that both recurrent and feedforward networks … can approximate any input-output function to any degree of accuracy," and on characterizing some ToCs as holding "that recurrent processing is necessary and sufficient for consciousness" (6). But recurrent processing theory (RPT; Lamme 2006) is best understand not as computational, but as claiming that recurrent processing is, in various organisms, implemented by biological factors that result in consciousness. RPT does not hold that recurrent processing, construed purely computationally, results in consciousness. The unfolding argument applies less widely than DSH urge.

I discuss the third criterion below. But the fourth criterion, that a "ToC should be able to determine which systems, apart from awake humans, are conscious" (7), is surprising, since we currently have no way to evaluate what a ToC says about that.

Many nonhuman creatures are sometimes conscious by being awake, but absent an account of what such consciousness consists in, we can't apply that to other systems. And there's little agreement about whether the contents of those creatures are ever conscious, and no way now to settle that. We can hope to extend an independently successful ToC to adjudicate those unclear cases, but we must first evaluate ToCs by cases we already understand.

III. Higher-order-thought theory (HOTT) is, in every version, a theory of what it is for psychological contents to be conscious. It does not address what it is for systems to be conscious. So HOTT applies only to systems with psychological contents. Since it's not credible that "networks with fewer than ten neurons" (6) subserve psychological contents, the small-systems argument of the third criterion is irrelevant. This holds also for global-workspace theory (Dehaene and Naccache, 2001). And the large-systems argument appeals to unity, which DSH don't clarify.

DSH misconstrue HOTT as a theory of systems' being conscious: "HOTT is subject to the small network argument because any 2-stage computer program is conscious" (12). But it isn't credible that a system itself is conscious if it's not in a reasonable number of contentful states. Applying the small-systems argument also requires construing HOTT as computational, which holds of no version I'm aware of. Here, as with RPT above, DSH sometimes adapt ToCs to fit their criteria, something a focus on detail would constrain.

DSH acknowledge that HOTT would not be purely computational if, as I've argued (Rosenthal 2005), higher-order thoughts are something like ordinary thinking. But then, they urge, the other-systems criterion would require HOTT "to explain what is crucial to be a 'thought' and which systems can have equivalents to 'thoughts' and be conscious, apart from humans" (12).

But it's not the job of a ToC to explain what thoughts are. Even when contents are conscious, we need an independent theory to explain what contents are (as in Rosenthal 2005, chs. 3, 7, 10). Indeed, divergent ToCs might rely on the same theory of content. A focus on conscious systems wrongly encourages seeing ToCs as addressing every aspect of conscious psychological reality.

The computational character of the second and third criteria and the demand that a ToC pronounce on systems of any type suggest a picture of consciousness as independent of the specific nature of implementing systems. It's arguably more promising to see consciousness as tightly tied to the relevant functioning of any such system.

REFERENCES

Dehaene, Stanislas and Lionel Naccache (2001), "Towards a Cognitive Neuroscience of Consciousness: Basic Evidence and a Workspace Framework," Cognition 79, 1-2 (April): 1-37. https://doi.org/10.1016/S0010-0277(00)00123-2

Doerig, Adrien, Aaron Schurger, and Michael H. Herzog (forthcoming), "Hard Criteria for Empirical Theories of Consciousness," Cognitive Neuroscience. https://doi.org/10.1080/17588928.2020.1772214

Lamme, Victor A. F. (2006), "Towards a True Neural Stance on Consciousness," Trends in Cognitive Sciences, 10, 11 (November): 494–501. https://doi.org/doi:10.1016/j.tics.2006.09.001

Rosenthal, David (2005), Consciousness and Mind, Oxford: Clarendon Press.