Stability is not intrinsic

D. C. Dennett and C. F. Westbury
Center for Cognitive Studies, Tufts University, Medford, MA 02144.
ddennett@tufts.edu cwestbur@emerald.tufts.edu
www.tufts.edu/as/cogstud/mainpg.htm

Abstract: A pure vehicle theory of the contents of consciousness is not possible. While it is true that hard-wired tacit representations are insufficient as content vehicles, not all tacit representations are hardwired. O'Brien & Opie's definition of stability for patterns of neural activation is not well-motivated and too simplistic. We disagree in particular with the assumption that stability in a network is purely intrinsic to that network. Many complex forms of stability in a network are apparent only when interpreted by something external to that network. The requirement for interpretation introduces a necessary functional element into the theory of the contents of consciousness, suggesting that a pure vehicle theory of those contents will not succeed.

One can be grateful for a theory such as the one offered, without being convinced by it, since O'Brien & Opie (O&O) resolutely explore some tempting but foggy territory. If our verdict about their exploration is negative, at least now we may be able to see clearly for the first time why we were wise to sidestep this option.

O&O's criticisms of the prevailing assumptions about unconscious information processing are timely and important. Although we have some minor quarrels with some of them, we agree that the standard assumption that there is a sharp (or principled) distinction between unconscious and conscious information-processing is misbegotten. They say: "it is not unreasonable to reserve judgment concerning the dissociability of explicit mental representation and phenomenal experience" (sect. 2.4, para. 4). We would put it somewhat more strongly. This oft presupposed dissociability depends on distinguishing between unconscious information processing on the one hand and very brief intervals of conscious-but-soon-forgotten information processing on the other, and this is not supportable. It presupposes what Dennett (1998) has called the myth of double transduction: the idea that unconscious contents in the brain become conscious by being transduced into a privileged neural medium (as most clearly expressed by Mangan 1993a; 1996).

The well-named "classical" approaches to cognitive science (whose name hints that they belong behind glass in a museum somewhere) do indeed propel the theorist headlong towards the myth of double transduction, but it is not clear that a pure vehicle theory can avoid equally ominous impasses in other directions. We see three related problems. The first concerns a missing taxon in O&O's representational taxonomy, the second their definition of stability, and the third the role that stability of component networks might play in a larger meta-network.

Transient tacit representations: As O&O point out, "hardwired" tacit representations can hardly serve the purposes of content vehicles in any theory of the fleeting contents of consciousness. However, they do not consider the question of whether all tacit representations are hardwired. They are not. Dennett's taxonomy

of styles of mental representations includes one further taxon which they overlook, transient tacit representations (Dennett, 1982, p. 224, reprinted in Dennett, 1987, pp. 213–25), which are available for a system's use only when that system is in a particular state. These representations are obviously the most important for the purposes of the argument presented. Indeed, the stable connectionist patterns championed by O&O are presumably just such sorts of mental representations – they call them non-explicit. Although O&O claim that the distinction between potentially explicit and tacit lapses "for all practical purposes," they are thinking only of hardwired, nontransient tacit representations. With transient tacit representations, that distinction is not simply of practical insignificance, but theoretically unmotivated.

The definition of stability: the idea that it is the most influential transient representations in cortical networks that earn the status of consciousness is fine. However, we do not see that O&O have succeeded in defining stability or its influence on the larger cortical network in such a way that one can assess their claim that "only stable patterns of activation are capable of encoding information in an explicit fashion in PDP systems" (sect. 5.1, para. 4); hence we also cannot assess their claim that it is all and only these stable patterns that are vehicles of conscious content.

One problem is simply that it is arbitrary and simplistic to declare that a network is stable if its constituent neurons are firing simultaneously and at a constant rate. Such a simple definition of stability ignores the fact that stability can manifest itself in a network in many more complex ways. Since a network can cycle through time, it can have a (possibly very complex) temporal stability that is impossible to discern spatially because it has no simple spatial representation at shorter time scales than the time it takes to cycle. Such complex stability can be discerned by an entity (including another network) which samples it at the right location and frequency. This idea of complex forms of stability was suggested by Hebb (1949) when he first described his Hebbian cell assembly, which is precisely the mechanism being described in this paper as the holder of phenomenal experience.

A further complication is added if we grant that the sampling system might have the ability to quantize states in the sampled system – that is, to pull information to its nearest category, as a basin of attraction in a complex system equates a wide number of states by the fact that they all lead to the same attractor. It is easy to imagine a network sampling a number of arbitrary points from another network and finding them stable because of its (the sampler's) characteristics, even though there is nothing in the sampled state that shows the stability. Stability is as much a function of the sampler as of the sampled. In a complex system, states that are not empirically identical can be functionally identical. We doubt that defining stability as simultaneous, constant firing will suffice to explain the behavior of myriads of interacting networks in the brain, and we are baffled by the suggestion that stability of the requisite sort is not to be found in serial simulations of connectionist networks – as if the stability of a virtual machine were any less powerful a feature than the stability of an actual machine.

The role of stability: finally, O&O's claim that it is a virtue of their vehicle theory that it makes phenomenal experience an "intrinsic, physical intranetwork property of the brain's neural networks" (sect. 5.1, para. 10) is, we think, confused. If the "intrinsic" property of stability is also an "intranetwork property," then presumably it is the role of the component networks in modulating the larger activities of the entire cortical metanetwork that mark them for the role of phenomenal experience, not their "intrinsic" stability. If it turned out, for instance, that there was a subclass of stable patterns in the networks that did not play any discernible role in guiding or informing potential behavior, would their stability alone guarantee their status as part of phenomenal experience? Why?

Dennett (1991) stressed the importance of this when he proposed what he called "the Hard Question" (p. 255): "and then what happens?" (see also And then What Happens?, Dennett, 1991, pp. 263–75). An instance of the failure to appreciate this

point appears in O'Brien & Opie's suggestion that "when phenomenal properties coincide temporally, . . . this is a consequence of the simultaneity of their vehicles" (sect. 5.3, para. 6). The "intrinsic" simultaneity of vehicles could not by itself account for subjective simultaneity. As we have stressed above, what matters is not actual ("intrinsic") simultaneity, but either the (correct or mistaken) detection of simultaneity by the larger system of which these vehicles are a part, or else the failure of the larger system to generate any complaint about their nonsimultaneity. If such functional effects are as vital as we suggest, a pure vehicle theory of consciousness cannot succeed.