The metaphysics of embodiment*

Shimon Edelman
Dept. of Psychology, Cornell University, Ithaca, NY 14853, USA

December 8, 2010

Abstract

Shanahan's eloquently argued version of the global workspace theory fits well into the emerging understanding of consciousness as a computational phenomenon. His disinclination toward metaphysics notwithstanding, Shanahan's book can also be seen as supportive of a particular metaphysical stance on consciousness — the computational identity theory.

In *Embodiment and the Inner Life*, Shanahan (2010) sets out to introduce and motivate a comprehensive approach to the understanding of consciousness, in the intuitive, pre-analytical sense of the word, as in "Having written the opening sentence of this review, I am now consciously weighing my options with regard to the directions in which I could make it go." Shanahan's self-avowed aversion to metaphysics makes him adopt a methodological stance similar to that of Crick and Koch (1990), whose bid to understand the neurobiology of consciousness famously limited itself to asking about the difference between those neurons whose activity correlate with the subject's phenomenal report and those whose activity does not. Accordingly, instead of aiming at a plausible definition of consciousness, Shanahan focuses on identifying the differences between conscious and unconscious modes of perceiving, thinking, and acting.

The book's approach is commendably comprehensive because it deals with this question on multiple levels: functional, computational, and neural-implementational. On the functional level, where the issues at stake have to do with evolutionary reasons and behavioral utility, Shanahan cites the mind's role in the organism's survival and reproduction and posits that the conscious mode of operation has evolved as a particularly effective way of dealing with emergencies. For much of the time, relatively compartmentalized, automatic processing suffices, but if something unexpected happens or if the going just gets tough, the whole of the embodied mind gets mobilized and is brought to bear on the problem at hand.

This view, which nicely engages the twin questions of why not all of our mental life is conscious at all times and why sometimes some of it is, is rooted in a long tradition in psychology. As Smith et al. (2003, p.338) put it, "If you watch an aging cat consider a doubtful leap onto the dryer, you will suspect that what James (1890, p.93) said is true, 'Where indecision is great, as before a dangerous leap, consciousness is agonizingly intense."' More generally, it is compatible with the idea that the mind is a tool for predicting the world (Dewey, 1910; Craik, 1943; Llinás, 2001; Bar, 2007; Edelman, 2008), which in turn affords informed planning for action.

*This manuscript is part of a collective review of "Embodiment and the Inner Life — Cognition and Consciousness in the Space of Possible Minds" by Murray Shanahan (2010), to appear in the *International Journal of Machine Consciousness*.

Bringing prediction explicitly into the explanatory picture (something that the book stops just short of doing) helps one understand not just the functional but also the computational and implementational layers of the proposed theory. Global mobilization is called for when prediction is likely to be particularly challenging. In such situations, the brain's global communications infrastructure "receives information from, and disseminates information to, numerous parallel processes operating on multiple levels, and thereby integrates their otherwise segregated activity. The integrative facility supplied by a global workspace gives rise to the conscious condition in general" (Shanahan, 2010, pp.68-69). The pooling of all available resources does not just increase the chances of predictive success: only a unified approach to prediction is ultimately warranted. For a cat, aging or not, "it would be most disadvantageous for the head to predict one thing and the tail to predict another" (Llinás and Roy, 2009, p.1305).

I perceive the theory put forward by Shanahan as complementing the analytical account of consciousness developed independently by Metzinger (2003) and Merker (2007) (for a synthesis, see Edelman, 2008, ch.9). On the M&M account, conscious experience is defined by a number of functional attributes: perspectivalness, agency, ownership, and selfhood. For each of these, a computational theory has been offered and a likely neurobiological substrate identified (importantly, the latter involves vertebrate-standard subcortical structures such as the superior colliculus; Merker, 2007).

According to M&M, the key functional component of the conscious state is a self-model — a virtual entity that the system constructs and puts in charge of guiding behavior, so as to concentrate the more important strands of decision making in a single computational bottleneck. Metzinger (2003, p.558) describes his model as a "total flight simulator," in which both the world and the pilot are virtual: "the brain, the dynamical, self-organizing system as a whole, *activates* the pilot if and only if it needs the pilot as a representational instrument in order to integrate, monitor, predict, and remember its own activities. As long as the pilot is needed to navigate the world, the puppet shadow dances on the wall of the neurophenomenological caveman's phenomenal state space. As soon as the system does not need a globally available self-model, it simply turns it off. Together with the model the conscious experience of selfhood disappears. Sleep is the little brother of death."

It is easy to imagine how the mechanism posited by Shanahan can provide Metzinger's virtual pilot (whose seat is in the superior colliculus; Merker, 2007) with integrated access to the entire wealth of informational patterns gleaned from past experience and encoded in the cortex (Merker, 2004). But why does such integrated access necessarily distinguish cognition that is phenomenally conscious from cognition that is unconscious? Shanahan grapples with this question on p.112: "How does integration [...] relate to phenomenology? The essential insight that answers each of these questions is this. Perfect integration occurs when the *being as a whole* is brought to bear on the ongoing situation."

One may object that this insight, while entirely valid, deflects the phenomenological question instead of answering it. Of course, this methodological move is fully in line with Shanahan's attitude toward metaphysics: "What does the present claim [...] tell us about phenomenology? To begin with, we must reject the overly metaphysical conception of consciousness that insists on a precisely definable content to a subject's consciousness at any given time, and that the question of content always has an answer. Instead, we must accept that there is something of the refrigerator light illusion about our inner lives. [...] If anything can be instructively said to be 'constitutive' of the conscious condition, it is the means by which the illusion is realized, the mechanism that switches on the light whenever the refrigerator door is opened, so to speak" (pp.114-115).

Given that a repudiation of metaphysics is in itself a metaphysical stance, I do not see what advantage it has over putting one's foot down and getting metaphysics to do some explanatory legwork for you (under strict supervision of science, of course). Warren McCulloch (1965) described just such a course of action in a paper titled *Through the den of the metaphysician*:

> Maxwell, who wanted nothing more than to know the relation between thoughts and the molecular motions of the brain, cut short his query with the memorable phrase, "but does not the way to it lie through the very den of the metaphysician, strewn with the bones of former explorers and abhorred by every man of science?" Let us peacefully answer the first half of his question "Yes," the second half "No," and then proceed serenely.

The metaphysical stance on phenomenal experience that I personally favor is computational identity. Given that the mind is fundamentally computational (Edelman, 2008), so is experience. This implies that experience is just as multiply realizable as, say, reasoning, and so is available in principle to computing machines other than biologically embodied brains. What makes a machine, biological or not, capable of experience? Because phenomenal experience cannot be a matter of attribution by an external observer, the relevant criterion must be intrinsic to the system in question. The Information Integration Theory (Tononi, 2008), which, as Shanahan notes, is close in spirit to his version of the global workspace theory, is one candidate framework within which an intrinsic account of experience can be sought.

Another theoretical framework that fits the prior requirements is based on intrinsic topological properties of the system's activity-space trajectory dynamics (Fekete and Edelman, 2010). A representational system whose dynamics gives rise to a parcellation of the space of its possible trajectories (so that certain regions of the system's activity space become excluded) constitutes a ready substrate for discernment, or categorization — arguably, the most fundamental property of experience. As the dynamics of such a system unfolds over time, the internal constraints that impose structure on the space of its possible trajectories get a chance to take effect. As Fekete and Edelman (2010) argue, this imbues the system's trajectories (but not instantaneous states) with intrinsic, system-internal meaning, singling them out as a possible vehicle of experience.[1]

It makes sense to go all the way here and *identify* temporally extended trajectories through a complex representation space with experience. In science, equating a formally defined entity with one that is merely intuitively described has precedents, such as the Church-Turing Thesis (Copeland, 2002), which declares effective computability (an intuitive notion) to be the same as Turing computability (a formal one). Resorting to a more mundane example, the metaphysical identity stance on conscious experience that I profess is akin to identifying dance with the ensemble of dancers in motion. What else could dance *be*?

Refusing to commit in this matter — as in saying instead "I don't know what dance is, but here's how you can tell if the event that's unfolding in front of you is *it*" — seems both unreasonable and counterproductive. We may guess where such reluctance comes from. Resistance to the metaphysical identity stance on the part of many brain/mind scientists, which often stems from their general aversion to metaphysics, is an

[1] The insistence of Fekete and Edelman (2010) that trajectories, but not time-frozen states, can serve as vehicles of experience, and that experience therefore must be inherently extended in time, seems compatible with Shanahan's postulate of the centrality of global integration to conscious experience: "To offer an integrated response, wherein the whole system is brought to bear on the ongoing situation, all potentially relevant processes in the system must be subject to the influence of that situation, and the system's response to it must take account of the activity of all potentially relevant processes. [...] System-wide influence is a property that can only be attributed retrospectively. Time has to elapse before any potential influence can be realized, and if events intervene it might never be realized" (Shanahan, 2010, p.113).

unfortunate legacy of the historical tendency of the latter to play the role of Spanish Inquisition to science's free thinking. In the words of McCulloch (1965),

> Our adventure is actually a great heresy. We are about to conceive of the knower as a computing machine. That is not a new heresy. It has already been prejudged by Dryden in *The Hind and the Panther*, when he says
>
> > And if they think at all, 'tis sure no higher
> > Than matter, set in motion, may aspire.
>
> I believe that he is correct, but I am not sure that that may not be high enough.

For my part, I am quite sure that the ongoing dynamics of a properly structured complex of activity-space trajectories is all there is to experience, and that therefore my inner life *is* the activity of my brain (cf. Metzinger, 2003, pp.58-59; Spivey, 2006, p.305; Edelman, 2008, p.488). Shanahan's book provides a useful characterization of the computational properties of this activity that make the realized experience more poignant — conscious — or less so.

References

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences 11*, 280–289.

Copeland, B. J. (2002). The Church-Turing Thesis. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*.

Craik, K. J. W. (1943). *The nature of explanation*. Cambridge, England: Cambridge University Press.

Crick, F. and C. Koch (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences 2*, 263–275.

Dewey, J. (1910). *How we think*. Lexington, MA: D. C. Heath.

Edelman, S. (2008). *Computing the mind: how the mind really works*. New York: Oxford University Press.

Fekete, T. and S. Edelman (2010). Prolegomena to a computational theory of experience. *Consciousness and Cognition*. Under review.

James, W. (1890). *The Principles of Psychology*. New York: Holt. Available online at http://psychclassics.yorku.ca/James/Principles/.

Llinás, R. and S. Roy (2009). The 'prediction imperative' as the basis for self-awareness. *Phil. Trans. R. Soc. Lond. B 364*, 1301–1307.

Llinás, R. R. (2001). *I of the Vortex*. Cambridge, MA: MIT Press.

McCulloch, W. S. (1965). Through the den of the metaphysician. In *Embodiments of mind*, pp. 142–156. Cambridge, MA: MIT Press. First published as "Dans l'antre du métaphysicien" in *Thales* 7:35-39 (1951).

Merker, B. (2004). Cortex, countercurrent context, and dimensional integration of lifetime memory. *Cortex 40*, 559–576.

Merker, B. (2007). Consciousness without a cerebral cortex: a challenge for neuroscience and medicine. *Behavioral and Brain Sciences 30*, 63–81.

Metzinger, T. (2003). *Being No One: The Self-Model Theory of Subjectivity*. Cambridge, MA: MIT Press.

Shanahan, M. (2010). *Embodiment and the Inner Life*. New York, NY: Oxford University Press.

Smith, J. D., W. E. Shields, and D. A. Washburn (2003). The comparative psychology of uncertainty monitoring and metacognition. *Behavioral and Brain Sciences 26*, 317–373.

Spivey, M. J. (2006). *The continuity of mind*. New York: Oxford University Press.

Tononi, G. (2008). Consciousness as integrated information: a provisional manifesto. *Biol. Bull. 215*, 216–242.