Harnad, S. (2011) Zen and the Art of Explaining the Mind. *International Journal of Machine Consciousness (IJMC)* (forthcming). [Review of Shanahan M. (2010) *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds.* Oxford University Press.]

Zen and the Art of Explaining the Mind

Stevan Harnad
Canada Research Chair in Cognitive Sciences
Université du Québec à Montréal
http://www.crsc.uqam.ca
&
School of Electronics and Computer Science
University of Southampton
http://users.ecs.soton.ac.uk/harnad/

Abstract. *The "global workspace" model would explain our performance capacity if it could actually be shown to generate our performance capacity. (So far it is still just a promissory note.) That would solve the "easy" problem. But that still would not explain how and why it generates consciousness (if it does). That's a rather harder problem.*

There are two problems of consciousness – the so-called "easy" problem (E), which is to explain how and why conscious entities are able to do what they can do, and the "hard" problem (H), which is to explain how and why they are conscious. Strictly speaking, E is not a problem of consciousness at all; it is a problem of explaining performance capacity: the capacity to move, sense, recognize, recall, categorize, identify, manipulate, learn, deduce, plan, problem-solve, speak and understand. Most of these capacities are capacities of conscious entities, to be sure, and many of them are exercised while the entitities are conscious – so we say they are done "consciously." But as long as we are only addressing the question of how and why they are done, rather than how and why they are done consciously, we are only addressing E, not H (Harnad 2003).

It is not altogether clear which of these problems Murray Shanahan's *Embodiment and the Inner Life: Cognition and Consciousness in the Space of Possible Minds* is addressing:

> "Although the word 'consciousness' makes regular informal appearances throughout this book, it would not be proper to characterize our aim as 'explaining consciousness'. Rather than trying to explain an amorphous something or other that no one can define clearly in the first place, our initial explanatory target is a distinction, the conscious/unconscious distinction… [S]ome of what we do we do consciously, but some of what we do we do unconsciously. Our task is to understand the nature of these contrasts… [to] operationalize the conscious/unconscious distinction… and to begin to account for them scientifically."

Shanahan does indeed address the conscious/unconscious distinction, mostly through phenomenology, thought experiments, and analogies, but I would not say he "operationalizes" it – if by "operationalize" we mean finding a set of empirical manipulations and observations that can then stand in for consciousness. The observations and manipulations are *correlates* of consciousness (sometimes reliable, sometimes not), so they can give us a good idea of whether or not a capacity or performance is likely to be executed consciously, but they are not the same thing as consciousness itself, nor do they explain it.

But before we go on, let us challenge Shanahan's contention that consciousness is "an amorphous something or other that no one can define clearly in the first place." Here's a definition: *To be conscious is to feel*. To do something consciously is to feel you are doing it. To be unconscious is not to feel anything. To do something unconsciously is to do it without feeling you are doing it. In other words, the conscious/unconscious distinction is the felt/unfelt distinction – and although I can be unsure about whether or not I have been injured, there is nothing amorphous or unclear about whether it feels painful (if/when it does feel painful); I can be unsure about whether or not I was touched, but I know I felt a touch; unsure about whether I moved, but felt I moved; unsure about whether I moved deliberately, but felt I moved deliberately; unsure about whether I understood, but felt I understood; unsure about whether I could do X, but felt I could do X; unsure about how I was doing X, but felt I was doing X.

All of us know exactly what it means to feel, and that's exactly what it means to be conscious. We also know exactly what (we think) is lacking in a stone, or a toaster, or a (contemporary) robot when we say it is unconscious: It does not feel anything. And feeling *anything at all* is what consciousness, and the problem of consciousness, is about. It is not particularly about feeling this or that; nor feeling this rather than that; nor even feeling this and not feeling that. It is about feeling anything at all. Explain how and why anything feels anything at all and you have solved the hard problem of consciousness (H).[1] Explain how and why something *does* something, and you are making inroads on the "easy" problem (E). But even if you point out – and sort out – the operational "correlates," behavioral, neural and phenomenological, of felt and unfelt doings (the conscious/unconscious distinction), you have not explained anything about consciousness: We already knew we feel. The fact that we feel some things and not others, that we feel under some conditions and not others, that we feel we are doing some of the things we do and not others, that we feel we know how we are doing some of the

---

[1] Shanahan asks:

> "Are we looking for a theory… of consciousness in humans alone?… [I]t must surely encompass other animals. Certain birds... are capable of remarkably intelligent behaviour, even though their brains are organised quite differently from our own. The brain of an octopus, another cognitively precocious animal, is even more alien."

We are not even looking for a theory of consciousness in "cognitively precocious" animals alone. Explain how and why an amphioxus feels "ouch" – even if that's the only thing it ever feels – and you've explained consciousness. But leave that out and you've bypassed consciousness altogether.

things we do and not others – all this just increases the mystery of how and why we feel at all, rather than helping to dispel it. Doing – and whatever it takes to generate the doing – seems to be the only functional component at play, and the only one needed, causally. Feeling floats along, correlated, and feeling as if it were causal; but its causal role is opaque.

Shanahan's book tries to dispel the mystery by blaming it on "metaphysical tendencies" that Wittgenstein (and Zen) should help us to overcome. But there is nothing metaphysical about asking why and how some entities feel (and some don't), and why and how some inputs and outputs are felt, and some are not. The problem is not metaphysical, it is epistemic. It is a *causal explanation* that is lacking, not a satori that dispels the sense that something real and important is being left unexplained.

Shanahan's substitute for an explanation is an interpretation: The "global workspace" in which processes fight it out for execution would be a useful contribution toward solving the easy problem (E) of explaining performance capacity -- if empirical evidence were provided that the model actually generates performance capacity more successfully than rival models. But it seems to me that the book is mostly showing how existing neural and computational models and data can be *interpreted* as if they were global-workspace models (or "not incompatible with" global workspace models) rather than showing empirically the superior power of workspace over non-workspace models in generating and explaining our performance capacities.

Let us set that aside, however, and agree that future empirical work may indeed show that a class of models fitting objective criteria for being global-workspace models do out-perform their rivals. Let's even suppose that they will prove to scale all the way up to being able to power embedded, embodied robots that are not only Turing-indistinguishable in their performance capacities from any of the rest of us, but even capture the functional principles underlying the brain processes that generate those same capacities in us (Harnad & Scherzer 2008). The "easy" problem will be completely solved. But will we have learned anything about consciousness – i.e., about how and why some things are felt and some things are not?

On the face of it, one would think so, for the Turing-scale model would at least be able to predict what will be felt and what will not, on the basis of its correlates in the workspace component. It would even be able to explain the functional advantages of the integrative/competitive features of the model, in its successful generation of the performance capacity. But will it explain how and why those functional correlates of the states in which we normally feel are *felt*? For, on the face of it, their performance-generating powers would be identical if they were all unfelt. Like so many other attempted explanations of this sort, Shanahan's conflates, inextricably, (1) the objective performance-generating benefits of the successful functional components that are correlated with feeling with (2) the (unexplained) benefits of their being executed feelingly rather than merely being executed. This is simply a non-sequitur (and that is why it is just interpretation rather than explanation). It leaves feeling (consciousness) completely untouched, explanatorily speaking. The workspace-model would be identical (indeed Turing-instinguishable) for feelingless zombies. Yet we know we are not feelingless zombies; and let us even assume, for the sake of argument, that the

workspace-powered robot would not be a feelingless zombie either. The question still remains: how and why does it (or we) feel?

Shanahan will of course reply that I have not been sufficiently post-reflective and Zen to overcome the illusion created by the metaphysical tendency here: There is in fact no further question to ask.

Well, let me take advantage of the fact that Shanahan has summarized the gist of the model that he thinks dispels the need to ask the further question, by asking the question about it *in situ* at each critical point:

> "A complex environment affords an animal more possibilities for action than can be hard-wired into its brain... [C]ognition is inherently embodied insofar as its fundamental role is to modulate an animal's sensorimotor interaction with its environment... by discovering new possibilities for action, either by experiment or through imagination, and introducing them into the animal's repertoire. It follows... that an intimate link exists between cognition and consciousness. Specifically, the conscious condition facilitates the exploration of previously untried action combinations, which is especially beneficial in novel situations."

On the face of it, this sounds like there are a lot of possibilities, and the internal mechanism must test and find the right ones: How and why is that process "of exploration of previously untried action combinations" conscious (felt) rather than just executed, unconsciously (unfelt)?

> "Much of our waking lives is devoted to habitual, automatic behaviour, such as driving or cleaning our teeth. But the episodes in our lives that matter to us most are those that we can remember, that we can talk about, that we respond to emotionally, the episodes that engage us fully, in short the conscious episodes."

How and why is remembering, talking and responding conscious (felt) rather than just executed, unconsciously (unfelt)?

> "[T]he distinction between automatic behaviour and the conscious condition... [corresponds to] a contrast between localised brain activity and globally integrated neural states in which the whole brain, indeed the whole person (or animal), is brought to bear on the ongoing situation."

How and why are "globally integrated neural states" conscious (felt) rather than just executed, unconsciously (unfelt)?

> "How might the brain be organised so as to realise the globally integrated states that are... the hallmark of the conscious condition?... [T]he pattern of long-range neural connections that constitute the brain's communications infrastructure… enable[s] information and influence from around the brain to funnel into a connective core, from where it can be broadcast back out again...[T]his connective core... acts as a global neuronal workspace, a serial procession of thoughts... distilled from the activity of massively many parallel processes, and unity arises out of multiplicity."

How and why is this "distilled unity out of multiplicity" conscious (felt) rather than just executed, unconsciously (unfelt)?

> "[T]he electrical activity of the brain displays exquisitely patterned interacting rhythms. Among these patterns, episodes of synchronised activity can be discerned at multiple frequencies, across widely separated sites…[L]ong-distance synchronised activity is a signature of the conscious condition, indicating that a coalition of brain processes is co-operating and communicating via the global neuronal workspace (connective core), to the exclusion of rival coalitions."

How and why is this "synchronized long-distance coalition" conscious (felt) rather than just executed, unconsciously (unfelt)?

> "One hallmark of sophisticated cognition is the ability to respond to novelty by effectively recombining the elements of an established behavioural repertoire. In terms of neural dynamics, this amounts to the capacity to explore an open-ended repertoire of coalitions of distributed brain processes… [T]this is facilitated by the conscious condition, wherein new coalitions of brain processes can form thanks to the involvement of the global neuronal workspace, which allows channels of communication to open up between pairs of brain processes that are not already associated."

How and why is this "allowing channels of communication to open up between pairs of brain processes that are not already associated" conscious (felt) rather than just executed, unconsciously (unfelt)?

The trouble with hermeneutics (as Hamlet archly points out to Polonius) is that when you are merely interpreting, rather than explaining, anything can be interpreted as anything else. Nowhere is this metaphoric tendency stronger than in espying the counterparts of our consciousness in the innocent innards of a performance model. Solving neither H nor E will be that easy.

References

Harnad, S. (2003) Can a Machine Be Conscious? How? *Journal of Consciousness Studies* 10(4-5): 69-75. http://eprints.ecs.soton.ac.uk/7718/

Harnad, S. and Scherzer, P. (2008) First, Scale Up to the Robotic Turing Test, Then Worry About Feeling. *Artificial Intelligence in Medicine* 44(2): 83-89 http://eprints.ecs.soton.ac.uk/14430/