Harnad, S. (1995) Why and How We Are Not Zombies. *Journal of Consciousness Studies* **1**: 164-167. (Presented at Royal Society/Association of British Sciences Writers Press Conference on "Consciousness: Its Place in Contemporary Science" Tuesday 7 February 10am - 12:50 am at the Royal Society)

Why and How We Are Not Zombies

Stevan Harnad
Cognitive Sciences Centre
Department of Psychology
University of Southampton
Highfield, Southampton
SO17 1BJ UNITED KINGDOM
phone: +44 01703 592582
fax: +44 01703 594597
http://cogsci.ecs.soton.ac.uk/~harnad/

ABSTRACT: *A robot that is functionally indistinguishable from us may or may not be a mindless Zombie. There will never be any way to know, yet its functional principles will be as close as we can ever get to explaining the mind.*

Let us not mince words. The difference between something that is and is not conscious is that something's home in something that's conscious, something experiencing experiences, feeling feelings, perhaps even, though not necessarily, thinking thoughts. Don't be lured into details about "self-awareness" and "intentionality." If there's something home in there, something hurting when pinched, then that's a mind and we are faced squarely with two age-old philosophical problems:

The first is: How can we know whether or not something's home in there? We aren't mind-readers. Not even a brain surgeon can guarantee that a patient is conscious. This is called the "other-minds" problem, and it's important to note that it is unlike any other problem in science having to do with the existence or reality of something that is unobservable. Quarks, like consciousness, cannot be observed directly, but there are many things that follow from quarks' existing or not existing, and those things can be observed. Does anything follow from the existence of consciousness, that would not follow just as readily if we were all Zombies who merely acted exactly as if they were conscious?

Think about it: Zombies who acted exactly as if they were conscious: Acted for how long? Well, for a lifetime obviously. And what does "exactly" mean? It means that there is no way to tell them apart from one of us based on anything they do. Zombies are functionally equivalent to and functionally indistinguishable from ourselves.

"So cut them apart," you say, "and check what's inside. If it's different from what's in us, that's still an observable difference, and we could conclude from that that they were just unconscious Zombies."

But could we really draw that conclusion if they were made of different stuff? What if they came from another planet: Would the fact that their innards were different be enough to convince you that they didn't feel pain when they were pinched and screamed? Would you yourself like to submit to such a verdict on another planet?

Or would you feel more comfortable pronouncing such a verdict if they didn't come from another planet, but were built in a lab here on earth? Is there something about that that guarantees that their screams are not genuine? If you feel there is, then you must feel that you know something about the solution to the second philosophical problem, the mind/body problem:

What is consciousness? Let us assume that, whatever it is, it isn't an extra "force" in nature, on a par with electricity or gravity, for otherwise all our thoughts would be telekinetic, mind moving matter, and high energy psychic forces would be duelling with their "duals," high energy physics forces, not only in the world as a whole, but in the Academy in particular, with the prize being the truth or falsity of the laws of energy conservation and perhaps even causality itself.

So we will assume, instead, that consciousness is not an autonomous force, but some property or aspect of the ordinary physical forces we already know. If so, then it is incumbent on anyone who thinks he can

tell the Zombie from the real thing that he be able to say what this property is. This is a notoriously difficult thing to do; in fact, I'm willing to bet it's impossible, and will even say why:

Pick a property. Any property. It can be anatomical, physiological, chemical or even "functional." Suppose that that property is what determines whether or not something is conscious. Now answer the following two questions:

(1) How could you ever determine whether that supposition -- that that's the property that distinguishes conscious things from unconscious ones -- was correct? That's the other-minds problem again.

But now let's suppose that the supposition -- that that's the property that distinguishes conscious things from unconscious ones -- was, miraculously, true, even though there was no way we could know it was true:

(2) In what, specifically, would its truth consist? What is it that something would lack if it lacked consciousness yet had the property you picked out? For if you pick anything other than consciousness itself as the thing it would lack if it lacked that property that was supposed to be the determinant of consciousness (which would be a bit circular), then one can always say: why can't it have that property without the consciousness? And no one has even the faintest inkling of what could count as a satisfactory answer to that question.

Console yourself with the fact that you are not alone, in facing this problem. It's not just centuries of philosophers who have wrestled with it in vain (and don't let anyone tell you the problem's only as old as Descartes, or that it's Descartes' fault, or anything like that: the problem of mind is as old as philosophy and it besets anyone who reflects on the nature of the mind): In particular, it is not only neurosurgeons, experimental psychologists, and ordinary people who are not mind-readers: The Blind Watchmaker (Who designed us though trial and error based on random mutations and their consequences for survival and reproduction) is no mind-reader either. He could not have let the conscious ones through and exclude the Zombies, because the two are functionally equivalent and functionally indistinguishable, and survival and reproduction are purely functional matters!

So what's a scientist to do, if he makes the mistake of staking out the mind as his terrain of inquiry? If the other-minds and mind/body problem are insoluble, does that mean that the mind is not scientifically investigable?

Only that it cannot be investigated directly, the way most things are investigated. It can be investigated indirectly, however, and perhaps eventually cornered by a series of approximations. Consider that we have been pretty cavalier about the problem of designing a Zombie: Doing it is not as easy as imagining it. There are plenty of formidable scientific problems to solve before we need to begin worrying about whether or not the functionally equivalent Zombies we've designed are conscious: We first have to generate their functional capacities.

Actually, I think scientific mind-theory is better described as reverse bioengineering: Ordinary engineering applies basic physics and engineering principles to the design of systems with certain functional capacities that are useful to us [bridges, ovens, planes, computers], whereas a scientific theory of mind would first have to successful second-guess what gives creatures like us, already ready-made by the Blind Watchmaker, our functional capacities.

So the road ahead of us is pretty clear for the time being, even though we have reason to believe there is a cloud at the end of it. For now, we need to devote our time and ingenuity to second-guessing those functional capacities until we manage to scale up to a Zombie. It should be some consolation that the usual rules of scientific inquiry are in effect for the functional part of our quest. It's easy to reverse-engineer a few isolated pieces of our functional capacity, and there are many different ways to do it, but as the functional chunk we take on gets bigger and bigger, the number of different ways it can be successfully generated gets smaller and smaller.

This is ordinary scientific underdetermination: You can always predict and explain a small body of data in lots of ways, most or all of which have nothing to do with reality. But as you predict and explain more and more data, your degrees of freedom shrink and your theory gets more powerful and general. The hope, in all areas of science, is that when it is complete, and predicts and explains all observable data, then your theory will have converged on reality; it will be the true theory of the way things are. It might not be. Perhaps there will be another theory that explains it all too, and there won't be any way to know which one's true. (Even picking the simpler theory, if one of them is simpler than the other, may not be the right choice, because the world may simply not happen to be the simplest one it might have been,

while still preserving all appearances.)

This is very much the way I think it will be at the end of the day (or at the end of the road, rather, if we stick to our previous metaphor), when we have reverse-engineered a complete Zombie, functionally equivalent to and functionally indistinguishable from us in any way. There is of course the possibility that there will be several, radically different, but equally successful Zombie designs. Cutting them (and ourselves) up, at that point, may be the only remaining way to narrow down the differences. We could insist that in the case of the reverse engineering of the mind, "all the observable data" means not only all the behavioral data, but all the neural data too, and we may want to put our money only on the Zombie that is indistinguishable from us in both respects.

I somehow doubt that will be necessary though. I really think that the task of generating our full Zombie capacity probably already narrows the degrees of freedom enough to exclude all nonconscious candidates. I draw some solace, for example, from the fact to which I have already drawn attention, namely, that the "forward engineer" (the Blind Watchmaker) whose work we are reverse engineering had nothing stronger to go by either. But does this mean that the mind/body problem is really just another example of scientific underdermination that will be settled by whatever candidate makes it to the home stretch at the end of the day?

Not quite. For that would be all there was to it if consciousness were like quarks, that other example of an unobservable that I mentioned earlier. One can, without too much loss of sleep, accept that if the winning theory says there are quarks -- because with quarks it can predict and explain all the observable data, whereas without them it can't -- then it's safe to accept that there are indeed quarks.

But I have to remind you that our complete reverse engineering theory, the one that generates our full Zombie capacity, will be entirely mute about consciousness, and will be just as capable of predicting and explaining all the observable data with or without the supposition that the Zombie is conscious.

Perhaps another way of putting it is that the complete Zombie theory will explain all the data except one: The fact of the existence of consciousness itself. This fact is at the heart (or rather the mind) of the very idea of "observation," and it's a fact that each of us can "observe" to be true in his own particular case.

So clearly the Zombie theory has left something out. Hence there is still something different here, something special about the mind/body problem, and something that eludes a scientific theory of mind unlike anything analogous in a scientific theory of matter. Maybe it's safe to assume that consciousness will somehow piggyback on Zombie capacity; maybe not. It might be some consolation that if it doesn't, we can never hope to be the wiser. But I think it's nothing to lose sleep about, at least not for a long time to come.

Harnad, S. (1982) Consciousness: An afterthought. *Cognition and Brain Theory* 5: 29 - 47.

Harnad, S. (1984) What are the scope and limits of radical behaviorist theory? *Behavioral and Brain Sciences* 7: 720 -721.

Harnad, S. (1989) Minds, Machines and Searle. *Journal of Theoretical and Experimental Artificial Intelligence* 1: 5-25.

Harnad, S. (1990) The Symbol Grounding Problem. *Physica D* 42: 335-346. [Reprinted in Hungarian Translation as "A Szimbolum-Lehorgonyzas Problemaja." *Magyar Pszichologiai Szemle* XLVIII-XLIX (32-33) 5-6: 365-383.]

Harnad, S. (1991) Other bodies, Other minds: A machine incarnation of an old philosophical problem. *Minds and Machines* 1: 43-54.

Harnad, S. (1992) Connecting Object to Symbol in Modeling Cognition. In: A. Clarke and R. Lutz (Eds) *Connectionism in Context*. Springer Verlag.

Harnad, S. (1992) The Turing Test Is Not A Trick: Turing Indistinguishability Is A Scientific Criterion. *SIGART Bulletin* 3(4) (October) 9 - 10.

Hayes, P., Harnad, S., Perlis, D. & Block, N. (1992) Virtual Symposium on Virtual Mind. *Minds and Machines* 2: 217-238.

Harnad, S. (1993) Grounding Symbols in the Analog World with Neural Nets.

Think 2(1) 12 - 78 (Special issue on "Connectionism versus Symbolism," D.M.W. Powers & P.A. Flach, eds.). [Also reprinted in French translation as: "L'Ancrage des Symboles dans le Monde Analogique a l'aide de Reseaux Neuronaux: un Modele Hybride." In: Rialle V. et Payette D. (Eds) La Modelisation. LEKTON, Vol IV, No 2.]

Harnad, S. (1993) Artificial Life: Synthetic Versus Virtual. Artificial Life III. *Proceedings, Santa Fe Institute Studies in the Sciences of Complexity. Volume XVI*.

Harnad, S. (1993) Symbol Grounding is an Empirical Problem: Neural Nets are Just a Candidate Component. Proceedings of the Fifteenth Annual Meeting of the Cognitive Science Society. NJ: Erlbaum

Harnad, S. (1993) Problems, Problems: The Frame Problem as a Symptom of the Symbol Grounding Problem. *PSYCOLOQUY* 4(34) frame-problem.11.

Harnad S. (1993) Discussion (passim) In: Bock, G.R. & Marsh, J. (Eds.) *Experimental and Theoretical Studies of Consciousness*. CIBA Foundation Symposium 174. Chichester: Wiley

Harnad, S. (1993) Turing Indistinguishability and the Blind Watchmaker. Presented at Conference on "Evolution and the Human Sciences" London School of Economics Centre for the Philosophy of the Natural and Social Sciences 24 - 26 June 1993.

Harnad, S. (1993) Grounding Symbolic Capacity in Robotic Capacity. In: Steels, L. and R. Brooks (eds.) *The "artificial life" route to "artificial intelligence." Building Situated Embodied Agents*. New Haven: Lawrence Erlbaum

Harnad, S, (1994) Does the Mind Piggy-Back on Robotic and Symbolic Capacity? To appear in: H. Morowitz (ed.) *The Mind, the Brain, and Complex Adaptive Systems*.

Harnad, S. (1994) Levels of Functional Equivalence in Reverse Bioengineering: The Darwinian Turing Test for Artificial Life. *Artificial Life* 1(3): 293-301.

Harnad S, (1994) The Convergence Argument in Mind-Modelling: Scaling Up from Toyland to the Total Turing Test. *Cognoscenti* 1:

Harnad, S. (1994) Computation Is Just Interpretable Symbol Manipulation: Cognition Isn't. Special Issue on "What Is Computation" *Minds and Machines* (in press)