Harnad, S. (2011) Deceiving Ourselves About Self-Deception. [Commentary of Von Hippel & Trivers "The Evolution and Psychology of Self-Deception"] *Behavioral and Brain Sciences* 34:25-26

DECEIVING OURSELVES ABOUT SELF-DECEPTION

Commentary on Von Hippel & Trivers on Self-Deception, BBS, 2010

Stevan Harnad
Canada Research Chair in Cognitive Sciences
Université du Québec à Montréal
&
School of Electronics and Computer Science
University of Southampton

Abstract: Were we just the Darwinian adaptive survival/reproduction machines Von Hippel & Trivers invoke to explain us, the self-deception problem would not only be simpler but non-existent. Why would *unconscious* robots bother to misinform *themselves* so as to misinform others more effectively? But as we are indeed conscious rather than unconscious robots, the problem is explaining the causal role of consciousness itself, not just its supererogatory tendency to misinform itself so as to misinform (or perform) better.

Von Hippel & Trivers (VH & T) are deceiving themselves -- with the help of adaptivist psychodynamics and a Darwinian Unconscious. They have not proposed an adaptive function for self-deception; they have merely clad adaptive interpersonal behaviour in a non-explanatory mentalistic interpretation:

I can persuade you more convincingly that I am unafraid of you (or better fool you into thinking that the treasure is on the right rather than the left, where it really is) if I am unaware of -- or "forget" -- my own fear (or the fact that the treasure is really on the left rather than the right).

Sure. But then in what sense am I afraid at all (or aware where the treasure really is)? If I *feel* (hence act) afraid, then you detect it. If I don't feel the fear (or the sinistroversive urge), then I don't act afraid, and you don't detect any fear (because there's nothing there to detect).

So in what sense am I "self-deceived"? (Ditto for left/right.) Is it always self-deception not to feel afraid (or not to remember that the treasure's on the right), when I "ought to" (or used to)?

The same is true of "self-enhancement": Yes, I am more convincing to others, hence more influential on their behaviour, if I behave as if I expect to succeed (even when I have no objective grounds for the expectation). But in what sense am I self-deceived? In feeling brave and confident, when I "ought to" be feeling fearful and pessimistic? Shouldn't organisms all simply be behaving in such a way as to maximize their adaptive chances?

In fact, what does what organisms *feel* have to do with any of this at all (apart from the entirely unexplained fact that they do indeed feel, that their feelings are indeed correlated with their adaptive behaviour, and that their feelings do indeed feel causal to them)? The feelings themselves (i.e, consciousness) is much harder to situate in the adaptive causal explanation -- unless you believe in telekinesis (Harnad 2000)! (Hence I feel that VH & T have bitten off a lot more here, phenomenally, than they can ever hope to chew, functionally.)

The treasure is the best example of all, because that is about *facts* (data) rather than just feelings: Suppose I did indeed "know" at some point that the treasure was on the left -- in the sense that if at that point I could have reached for it without risk of being attacked by you, I would have reached for it on the left. But, according to VH & T, it was adaptive for me to "forget" where the treasure really was, and to believe (and behave as if) it was on the right rather than the left, so as to deceive you into heading off to the right so I could eventually grab the treasure on the left and dart off with it.

But isn't the true adaptive design problem for the Blind Watchmaker – apart from the untouched problem of how and why we feel at all (Harnad 1995) -- a lot simpler here than we are making it out to be (Harnad 2002)? And are we not deceiving ourselves when we "adapt" the adaptive explanation so as to square with our subjective experience?

All that's needed for adaptive cognition and behaviour is *information* (i.e., data). To be able to retrieve the treasure, what I (or rather my brain) must have have is reliable data on where the treasure really is -- on the left or the right. Likewise, in order to get you to head off toward the right, leaving the treasure to me, I need to be able to behave exactly as if I had information to the effect that it was on the right rather than the left (or as if I had no information at all). Adaptive "mind-reading" (*sensu* Premack & Woodruff 1978), after all, is just behavioral-intention-reading and information-possession-reading. It's not really telepathy.

Nor does it need to be. Insofar as the putative adaptive value of self-deception in interpersonal interactions is concerned, an adaptive behaviourist (who has foolishly -- and falsely -- deceived himself into denying the existence of consciousness) could easily explain every single one of VH & T's examples in terms of the adaptive value of mere deception -- behavioural deception -- of other organisms.

And when it comes to true "self-deception" -- do I really have to *forget* where the treasure actually is in order to successfully convince either you or me of something

that is adaptive for me? Well, there the only reason VH & T would seem to have a leg up on the blinkered adaptive behaviourist is that VH & T do *not* deceive themselves into denying consciousness (Harnad 2003). But what VH & T completely fail to do, is to explain (1) what causal role consciousness (feeling) itself performs in our adaptive success, let alone (2) what second-order causal role consciousness might need to perform in the kind of peekaboo game individual organisms sometimes seem to play with themselves. Both remain just as unexplained as they were before, and the first (1) is the harder problem, hence the one that needs to be solved first (Harnad & Scherzer 2008).

Might self-deception rather be a form of *anosognosia*, where our brains are busy making do with whatever informational and behavioural resources they have at their disposal, with no spare time to deceive us (inexplicably) into feeling that we're doing what we're doing because we *feel* like it?

Apart from that, it's simple to explain, adaptively, why people lie, cheat and steal (or try to overachieve, against the odds, or avoid untoward data): It's because it works, when it works. It's much harder to explain why we *don't* deceive, when we don't, than why we do, when we do. We usually distinguish between the sociopaths, who deceive without feeling (or showing) any qualms, and the rest of us. Have sociopaths deceived themselves about what's right and wrong, confusing true and false with being whatever it takes to get what one wants, whereas the rest of us are keeping the faith? Or are they just better Method Actors than the rest of us?

Harnad, Stevan (1995) "Why and How We Are Not Zombies. *Journal of Consciousness Studies* 1:164-167. http://cogprints.org/1601/

Harnad, S. (2000) Correlation vs. Causality: How/Why the Mind/Body Problem Is Hard. *Journal of Consciousness Studies* 7(4): 54-61. http://cogprints.org/1617/

Harnad, S. (2002) Turing Indistinguishability and the Blind Watchmaker. In: J. Fetzer (ed.) *Evolving Consciousness*. Amsterdam: John Benjamins. Pp. 3-18. http://cogprints.org/1615/

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/

Premack, D. & Woodruff, G. 1978. Does the chimpanzee have a theory of mind? *Behavioral & Brain Sciences* 1: 515-526.