Daniel C. Dennett

How Could I Be Wrong? How Wrong Could I Be?

One of the striking, even amusing, spectacles to be enjoyed at the many workshops and conferences on consciousness these days is the breathtaking overconfidence with which laypeople hold forth about the nature of consciousness — their own in particular, but everybody's by extrapolation. Everybody's an expert on consciousness, it seems, and it doesn't take any knowledge of experimental findings to secure the home truths these people enunciate with such conviction.

One of my goals over the years has been to shatter that complacency, and secure the scientific study of consciousness on a proper footing. *There is no proposition about one's own or anybody else's conscious experience that is immune to error, unlikely as that error might be.* I have come to suspect that refusal to accept this really quite bland denial of what would be miraculous if true lies behind most if not all the elaboration of fantastical doctrines about consciousness recently defended. This refusal fuels the arguments about the conceivability of zombies, the importance of a 'first-person' science of consciousness, 'intrinsic intentionality' and various other hastily erected roadblocks to progress in the science of consciousness.

You can't have infallibility about your own consciousness. Period. But you can get close — close enough to explain why it seems so powerfully as if you do. First of all, the intentional stance (Dennett, 1971; 1987) guarantees that any entity that is voluminously and reliably predictable as an intentional system will have a set of beliefs (including the most intimate beliefs about its personal experiences) that are mainly true. So each of us can be confident that *in general* what we believe about our conscious experiences will have an interpretation according to which we are, in the main, right. How wrong could I be? Not that wrong. Not about most things. There *has* to be a way of nudging the interpretation of your manifold beliefs about your experience so that it comes out largely innocent of error — though this might not be an interpretation you yourself would be inclined to endorse. This is not a metaphysical gift, a proof that we live in the best of all possible worlds. It is something that automatically falls out of the methodology: when adopting the intentional stance, one casts about for a maximally

Journal of Consciousness Studies, **9**, No. 5–6, 2002, pp. ??–??

charitable (truth-rendering) interpretation, and there is bound to be one if the entity in question is hale and hearty in its way.

But it does not follow from this happy fact that there is a path or method we can follow to isolate some privileged set of guaranteed-true beliefs. No matter how certain you are that *p*, it may turn out that *p* is one of those relatively rare errors of yours, an illusion, even if not a grand illusion. But we can get closer, too. Once you have an intentional system with a capacity for communicating in a natural language, it offers itself as a candidate for the rather special role of self-describer, not infallible but *incorrigible* in a limited way: it may be wrong, but there may be no way to correct it. There may be no truth-preserving interpretation of all of its expressed *opinions* (Dennett, 1978; 1991) about its mental life, but those expressed opinions may be the best source we *could* have about what it is like to be it. A version of this idea was made (in-)famous by Richard Rorty back in his earlier incarnation as an analytic philosopher, and has been defended by me more recently in *The Case for Rorts* (Dennett, 2000). There I argue that if, for instance, Cog, the humanoid robot being developed by Rodney Brooks and his colleagues at MIT, were ever to master English, its own declarations about its subjectivity would systematically tend to trump the 'third-person' opinions of its makers, even though they would be armed, in the limit, with perfect information about the micro-mechanical implementation of that subjectivity. This, too, falls out of the methodology of the intentional stance, which is the only way (I claim) to attribute content to the states of anything.

The price we pay for this near-infallibility is that our heterophenomenological worlds may have to be immersed in a bath of metaphor in order to come out mainly true. That is, our sincere avowals may have to be rather drastically reconstrued in order to come out literally true. For instance, when we sincerely tell our interrogators about the mental images we're manipulating, we may not *think* we're talking about convolutions of data-structures in our brain — we may well *think* we're talking about immaterial ectoplasmic composites, or intrinsic qualia, or quantum-perturbations in our micro-tubules! — but if the interrogators rudely override these ideological glosses and disclaimers of ours and forcibly re-interpret our propositions as actually being *about* such data-structure convolution, these propositions will turn out to be, in the main, almost all true, and moreover deeply informative about the ways we solve problems, think about the world, and fuel our subjective opinions in general. (In this regard, there is nothing special about the brain and its processes; if you tell the doctor that you have a certain sort of travelling pain in your gut, your doctor may well decide that you're actually talking about your appendix — whatever you may think you're talking about — and act accordingly.)

Since we are such reflective and reflexive creatures, we can participate in the adjustment of the attributions of our own beliefs, and a familiar philosophical 'move' turns out to be just such reflective self-re-adjustment, but not a useful one. Suppose you say you know just what beer tastes like to you now, and you are quite sure you remember what beer tasted like to you the first time you tasted it, and you can compare, you say, the way it tastes now to the way it tasted then.

Suppose you declare the taste to be the same. You are then asked: Does anything at all follow from this subjective similarity in the way of further, objectively detectable similarities? For instance, does this taste today have the same higher-order effects on you as it used to have? Does it make you as happy or as depressed, or does it enhance or diminish your capacity to discriminate colours, or retrieve synonyms or remember the names of your childhood friends or. . . . .? Or have your other, surrounding dispositions and habits changed so much in the interim that it is not to be expected that the very same taste (the same *quale,* one may venture to say, pretending to know what one is talking about) would have any of the same effects at this later date? You may very well express ignorance about all such implications. *All you know*, you declare, is that this beer now *tastes just like* that first beer did (at least in some ineffable, intrinsic regard) *whether or not* it has any of the same further effects or functions. But by explicitly jettisoning all such implications from your proposition, you manage to guarantee that it has been reduced to a vacuity. You have jealously guarded your infallibility by seeing to it that you've adjusted the content of your claim all the way down to zero. You can't be wrong, because there's nothing left to be right or wrong about.

This move is always available, but it availeth nought. It makes no difference, by the way, whether you said the beer tastes the same or different; the same point goes through if you insist it tastes different now. Once your declaration is stripped of all powers of implication, it is an empty assertion, a mere demonstration that *this* is how you fancy talking at this moment. Another version of this self-vacating move can be seen, somewhat more starkly, in a reaction some folks opt for when they have it demonstrated to them that their colour vision doesn't extend to the far peripheries of their visual fields: They declare that on the contrary, their colour *vision* in the sense of colour *experience* does indeed extend to the outer limits of their phenomenal fields; they just disavow any implications about what this colour experience they enjoy might enable them to do — for example, identify by name the colours of the objects there to be experienced! They are right, of course, that it *does not follow* from the proposition that one is having colour experiences that one can identify the colours thus experienced, or do better than chance in answering same-different? questions, or use colour differences to detect shapes (as in a colour-blindness test) to take the most obvious further effects. But if *nothing* follows from the claim that their peripheral field is experienced as collared, their purported disagreement with the researchers' claim that their peripheral field lacks colour altogether evaporates.

O'Regan and Noë (2001) argue that my heterophenomenology makes the mistake of convicting naive subjects of succumbing to a grand illusion.

> But is it true that normal perceivers think of their visual fields this way [as in sharp detail and uniform focus from the centre out to the periphery]? Do normal perceivers really make this error? We think not . . . normal perceivers do not have ideological commitments concerning the resolution of the visual field. Rather, they take the world to be solid, dense, detailed and present and they take themselves to be embedded in and thus to have access to the world. [p. XXX]

My response to this was:

> Then why do normal perceivers express such surprise when their attention is drawn to facts about the low resolution (and loss of colour vision, etc) of their visual peripheries? Surprise is a wonderful dependent variable, and should be used more often in experiments; it is easy to measure and is a telling betrayal of the subject's *having expected something else*. These expectations are, indeed, an overshooting of the proper expectations of a normally embedded perceiver–agent; people shouldn't have these expectations, but they do. People are shocked, incredulous, dismayed; they often laugh and shriek when I demonstrate the effects to them for the first time. (Dennett, 2001, p. XXXX)

O'Regan and Noë (see also Noë *et al.*, 2000; Noë, 2001; and Noë and O'Regan, forthcoming) are right that it need not seem to people that they have a detailed picture of the world in their heads. But typically it does. It also need not seem to them that they are not 'zombies' but typically it does. People *like* to have 'ideological commitments'. They are inveterate amateur theorisers about what is going on in their heads, and they can be mighty wrong when they set out on these paths.

For instance, quite a few theorisers are very, very sure that they have something that they sometimes call original intentionality. They are prepared to agree that interpretive adjustments can enhance the reliability of the so-called reports of the so-called content of the so-called mental states of a robot like Cog, because those internal states have only *derived* intentionality, but they are of the heartfelt opinion that we human beings, in contrast, have the real stuff: we are endowed with genuine mental states that have content quite independently of any such charitable scheme of interpretation. That's how it seems to them, but they are wrong.

How could they be wrong? They could be wrong about this because they could be wrong about anything — because they are not gods. How wrong could they be? Until we excuse them for their excesses and re-interpret their extravagant claims in the light of good third-person science, they can be utterly, bizarrely wrong. Once they relinquish their ill-considered grip on the myth of first-person authority and recognize that their limited incorrigibility depends on the liberal application of a principle of charity by third-person observers who know more than they do about what is going on in their own heads, they can become invaluable, irreplaceable informants in the investigation of human consciousness.

References

Dennett, D.C. (1971), 'Intentional systems', *J.Phil*, **68**, pp. 87–106.
Dennett, D.C. (1978), 'How to change your mind' in *Brainstorms* (Cambridge, MA: MIT Press).
Dennett, D.C. (1987), *The Intentional Stance* (Cambridge, MA: MIT Press).
Dennett, D.C. (1991), *Consciousness Explained* (Boston: Little, Brown, and London: Allen Lane, 1992).
Dennett, D.C. (2000), 'The Case for Rorts', in *Rorty and his Critics*, ed. Robert Brandom (Oxford: Blackwells).
Dennett, D.C. (2001), 'Surprise, surprise', commentary on O'Regan and Noë, 2001, *BBS*, **24** (5), pp. xxxx.
Noë, A., Pessoa, L. and Thompson, E. (2000), 'Beyond the grand illusion: what change blindness really teaches us about vision', *Visual Cognition*, **7**, pp. 93–106.
Noë, A. (2001), 'Experience and the active mind', *Synthese*, **129**, pp. 41–60.
Noë, A. and O'Regan, J.K. (Forthcoming), 'Perception, attention and the grand illusion',*Psyche*, **6** (15), URL: http://psyche.cs.monash.edu.au/v6/psyche-6-15-noe.html
O'Regan, J.K. and Noë, A. (2001), *Behavioral and Brain Studies*, **24** (5), pp.xxxxx.