Late night thoughts on "explaining" consciousness

Nicholas Humphrey, 2007, Unpublished Essay

In a PostScript to my book Seeing Red (which is on my website), I wrote:.

"One day someone will write a book that explains consciousness. The book will put forward a theory that closes the 'explanatory gap' between conscious experience and brain activity, by showing how a brain state *could in principle* amount to a state of consciousness. But it will do more. It will demonstrate just why this particular brain state *has to be* this particular experience. As Dan Lloyd puts it in his philosophical novel, *Radiant Cool*: 'What we need is a transparent theory. One that, once you get it, you see that anything built like *this* will have *this* particular conscious experience.'"

I thought when I wrote this -- referring approvingly to Lloyd, and going on to muse about whether anyone could yet provide such a "transparent theory" -- that the meaning of what I'll call "Lloyd's challenge" was obvious. But I've been worrying about it since, and have discovered that I for one really didn't understand what Lloyd was asking for.

Even if we simplify things considerably by assuming that what's at issue is *raw conscious sensation* (not, consciousness in any of several grander meanings of the term), I can see there are still a dozen unknowns and potential pitfalls in the way the challenge is being formulated. Problems indeed with almost every word, as I've highlighted them in the following version of it:

> What we need is a "transparent" "theory" that enables us, once we "get it", to "see" that "anything built like this" will "have" "this particular conscious sensation".

I now want to work through these problems step by step. I may seem to be being overly pedantic. But I hope you'll find it's worth it. Maybe if we can sort out what Lloyd's challenge can possibly amount to (if it amounts to anything) we may actually learn something crucial about what we can possibly mean by consciousness.

So, to start with, what might Lloyd mean by "theory"? And, beyond that, by "getting it" and "seeing" what it predicts?

It's clear, from what he goes on to demand of it, that what he calls a "theory" is actually more of an *explanation* or a *proof*. He's saying we need a form of argument as to why the occurrence of one thing, X, entails the existence of another, Y, such that once we understand how this argument goes (we "get it"), we can deduce (we can "see") that where there's X there will be Y.

Furthermore, since we are doing science and not pure logic or mathematics, he presumably means that this argument should conform to the canonical pattern of scientific explanation: that's to say, it should show why X *logically entails* Y, *given what else we know about how the world works*, i.e. the laws of nature.

That this must be the formal structure of the argument is, as we shall see, rather important to resolving several other issues..

To quote the Cambridge Dictionary of Philosophy, in general the aim of scientific explanation is to show how: "a statement describing the event to be explained is logically derived from the covering laws together with statements of antecedent conditions".

Suppose, for comparison, we want to explain why H2O has the properties of water (a case that is often used as an analogy for brains and consciousness). Then what our explanation must do is to show how a statement describing the properties of water can be logically derived from the laws of physics and chemistry together with a statement of the atomic structure of the water molecule.

And likewise, now, with consciousness, what the explanation has to do is to show how a statement describing something's having a particular sensation can be logically derived from certain laws together with a statement about how this something is built.

But then, the question arises, in the case of consciousness: just what are these *covering laws*?

Lloyd says we need a "transparent" theory. It's not clear exactly what he means by this. But I would assume what he means is that we should be able to see how the argument works from scratch, on the basis of *what we already know and understand*. In particular the theory must not *beg the question* by assuming what it's trying to prove. So, we should insist that the covering laws must be independently established..

This is a fairly standard constraint on what makes for a good explanation Just as we want our explanation of the wetness of water to flow from the laws of physics and chemistry, and not from a special science of wetness, so now we want our explanation of consciousness to flow from just those laws of nature that we already knew about prior to investigating consciousness.

It means we're not going to admit laws introduced late in the day to "save the appearances". No appeal to mysterious principles of emergence, such as, say, the claim that consciousness just pops up at a certain level of brain complexity. But nor, for that matter, an appeal to empirically established brute facts about how consciousness just happens to correlate with brain activity, such as the so-called "law of specific nerve energies" which states that as a matter of fact activity in different areas of sensory cortex just happens to give rise to different modalities of conscious sensation.

So, we can begin to add some clarification to Lloyd's challenge:

> What we need is a theory that enables us, once we understand the argument, to *work out by a combination of logic and independently established laws*, that anything built like this (the antecedent condition) will have this particular sensation (the thing to be explained).

But, next how about those phrases "anything built like this" and "this particular sensation"?

In both cases, presumably the term *"this"* is meant to serve an indexical role: we're *pointing* to something to define the thing in question: "*This* is the way the farmer rides". But what exactly is being pointed to? Is it meant to be an actual example, providing an ostensive definition: "*Here is an instance* of the way the farmer rides"? Or is it meant to be, rather, a

specification of the thing's key properties, providing a descriptive definition: "*Here, written on this notepad*, *is a description* of the way the farmer rides"?

Let's begin with "anything built like this". Given that the thing built like this is a brain of some sort (a real brain or an artificial one), are we going to define the brain in question by having it on display in front of us – in a brain-scanning machine, perhaps – and gesturing towards its peculiar features: *this here example of activity in this here bit of cortex*? Or are we, rather, going to define it by alluding to its key properties on a descriptive level, in words, in symbols, in diagrams : *the kind of brain state with this here physical / functional description*?

It's obvious, surely, that it has to be the latter. For one thing, it would certainly be an odd and unhelpful form of proof if the only people who could follow it were those who happened to be present to witness what the antecedent condition was. But, in any case, if we are going to argue *deductively* that where there's X there will be Y, we have to start with a *statement* about the occurrence of X – which can only mean a descriptive statement of some sort. It's obvious, if you think about it, that something defined ostensively simply can't feature as the premise of an argument, *only* something defined descriptively can.

So, now, we can add some further clarification:

> What we need is a theory that enables us, once we understand the argument, to work out by a combination of logic and independently established laws, that *anything built to meet this particular description of its physical/functional properties* will have this particular sensation.

Which brings us at last to what's meant by "this particular sensation"? There are those two possibilities again. Are we going to define the sensation in question by setting up a particular case of it – pricking our thumb, say, or sucking a lemon – and then identifying the sensation by a kind of inner gesture towards something happening in our own minds: *this here sensation I am having now* ? Or, here too, are we going to define it by alluding to its properties, as they play out at a publically describable/demonstrable level: *the kind of sensation which leads the subject to talk and believe and act in the ways here described?*

Hmm. Now, it's not so obvious that it has to be the latter. In fact I've no doubt most people's first thought when they hear the words "having this particular sensation" in this context, will be to go for an ostensive definition. That's to say, they'll focus in on something immediately present to them personally, and centre the discussion around that. So, for instance, they might indeed prick their thumb and think "Here am I experiencing what it's like to sense pain. Now, what I want the theory to do is to help me see just why this experience of *mine* is a consequence of my being built the way I am."

Maybe so. But, first thoughts aren't necessarily well informed thoughts.

For a start, since the indexical gesture towards the sensation is private and unverifiable, the only person who could possibly be in a position to see that the theory predicts that something built to meet such and such a description will have *this sensation* defined ostensively, is going to be the subject of the sensation as such – the one person in whose mind it's occurring. True, we might chose to suppose that actually, all human beings are built similarly and all have very similar experiences in the same circumstances – so that if you want to know what I'm pointing at when I prick my thumb and have a pain sensation you can create your own example of this pain by pricking your own thumb. Even so, we'd still be in the odd situation where only you and I (and anyone else who was prepared to join us in the ostensive demonstration) could understand the proof.

But, what's more, we're again going to have problems with the logic of the argument. If we are going to argue deductively that where there's X there will be Y, we can only end up with a *statement* about the occurrence of Y. Just as something defined ostensively can't feature as the premise of an argument, no more can it feature as the *conclusion.*

The upshot is that we really don't have a choice. Even though we *want* an ostensive definition, we'll be *obliged* to define the sensation descriptively. What we *want* is a theory that enables us to deduce that X will have *a sensation such as this example I'm pointing to in my own mind;* yet, all we're *allowed* is a theory that enables us to deduce that X will have *a sensation with this description*.

So, what to do? I think, when it comes to it, a compromise is possible (though it's actually more a case of caving in). We can agree to keep it personal by centring it on what we each know about ourselves – *my sensation*. But, like it or not, even though it's the particular sensation in our minds, we're going to have to go beyond ostension and describe the thing we're gesturing to inside our minds in propositional terms.

Thus, finally we've arrived at what I think Lloyd's challenge, all things considered, must boil down to:

> What we need is a theory that enables us, once we understand the argument, to work out by a combination of logic and independently established laws, that anything built to meet this description of its physical/functional properties *will have the kind of sensation that meets the description now before us that a first person subject will give of the kind of experience "I'm having now": viz.* "*a sensation, to which I have such and such attitudes, about which I have such and such beliefs, as to which I have such and such things to say (including claims about how marvellous , ineffable, intrinsically phenomenal it all is, etc)"*

And where's this get us with the more general issue of what it means for us to be conscious? I think that in discovering what we -- as philosophical theorists -- can and can't be seeking to explain when we try to explain consciousness, we've discovered something important about what we – as the subjects of consciousness -- actually have access to when we reflect on our experience.

Against all our deepest intuitions, the bottom line is that "*the phenomenal is propositional*." Dan Dennett rules! (Ah, but., there's a lot more to be said about just why sensations should have evolved *to appear to us to be* so phenomenally mysterious).

19/01/07