CANADIAN JOURNAL OF PHILOSOPHY
Volume 34, Number 2, June 2004, pp. 153-174

Reductive Explanation and the 'Explanatory Gap'

PETER CARRUTHERS
University of Maryland
College Park, MD 20742
USA

Can phenomenal consciousness be given a reductive natural explanation? Exponents of an 'explanatory gap' between physical, functional and intentional facts, on the one hand, and the facts of phenomenal consciousness, on the other, argue that there are reasons of principle why phenomenal consciousness cannot be reductively explained: Jackson (1982), (1986); Levine (1983), (1993), (2001); McGinn (1991); Sturgeon (1994), (2000); Chalmers (1996), (1999). Some of these writers claim that the existence of such a gap would warrant a belief in some form of ontological dualism (Jackson, 1982; Chalmers, 1996), whereas others argue that no such entailment holds (Levine, 1983; McGinn, 1991; Sturgeon, 1994). In the other main camp, there are people who argue that a reductive explanation of phenomenal consciousness is possible in principle (Block and Stalnaker, 1999), and yet others who claim, moreover, to have provided such an explanation in practice (Dennett, 1991; Dretske, 1995; Tye, 1995, 2000; Lycan, 1996; Carruthers, 2000). I shall have nothing to say about the ontological issue here (see Balog, 1999, for a recent critique of dualist arguments); nor shall I have a great deal to say about the success or otherwise of the various proposed reductive explanations. My focus will be on the explanatory gap itself — more specifically, on the question whether any such principled gap exists. I shall argue that it does not. The debate will revolve around the nature and demands of reductive explanation in general. And our focus will be on Chalmers and Jackson (2001) — hereafter 'C&J' — in particular, as the clearest, best articulated, case for an explanatory gap. While I shall not attempt to demonstrate this here, my view is that if the C&J argument can be undermined, then it will be a relatively straightforward matter to show that the other versions of the argument must fall similarly.

I Introduction: The Explanatory Gap

C&J argue as follows.

1. In the case of all macroscopic phenomena *M* not implicating phenomenal consciousness (and more generally, for all macroscopic phenomena *M* with the phenomenally conscious elements of *M* bracketed off), there will be an a priori conditional of the form, *(P & T & I) ⊃ M* — where *P* is a complete description of all micro-physical facts in the universe, *T* is a '*That's all*' clause intended to exclude the existence of anything not entailed by the physical facts, such as angels and non-physical ectoplasm, and *I* specifies indexically where *I* am in the world and when *now* is.

2. The existence of such a priori conditionals is required, if there are to be reductive explanations of the phenomena described on the right-hand sides of those conditionals.

3. So, if there are no a priori conditionals of the form, *(P & T & I) ⊃ C*, where *C* describes some phenomenally conscious fact or event, then it follows that phenomenal consciousness isn't reductively explicable.[1]

C&J indicate that Chalmers, though not Jackson, would make the further categorical claim that:

4. There are no a priori conditionals of the form, *(P & T & I) ⊃ C*.

Hence Chalmers, but not Jackson, would draw the further conclusion that phenomenal consciousness *isn't* reductively explicable.

I agree with Chalmers that premise (4) is true (or at least, true under one particular interpretation). I think we can see a priori that there is no a priori reducing conditional for phenomenal consciousness to be had, in the following sense. No matter how detailed a description we are given in physical, functional and/or intentional terms, it will always be

---

1 C&J actually present their case somewhat differently. They first argue that there is an a priori conditional of the form, *(P & T & I & C) ⊃ M*, where *C* is a description of all facts of phenomenal consciousness, and *M* includes *all* macroscopic facts. And they next argue subtractively, that if there is no a priori conditional of the form, *(P & T & I) ⊃ C*, then this must be because phenomenal consciousness isn't reductively explicable. Nothing significant is lost, and no questions are begged, by re-presenting the argument in the form that I have adopted in the text.

conceivable that those facts should be as they are, while the facts of phenomenal consciousness are different or absent, so long as those facts are represented using purely recognitional concepts of experience. We shall be able to think, 'There might be a creature of whom all *that* is true, but in whom *these* properties are absent,' where the indexical 'these' expresses a recognitional concept for some of the distinctive properties of a phenomenally conscious experience. So I accept that it will always be possible to conjoin any proposed reductive story with the absence of phenomenal consciousness, to form an epistemic / conceptual possibility. And I therefore also allow that some of the relevant conditionals, here, are never a priori — those conditionals taking the form $(P \,\&\, T \,\&\, I) \supset C$ (where C states the presence of some phenomenally conscious property, deploying a recognitional concept for it).

I shall be taking for granted, then, that we can possess purely recognitional concepts for aspects of our phenomenally conscious experience. (Arguments to the contrary from writers as diverse as Wittgenstein, 1953, and Fodor, 1998, are hereby set to one side.) This isn't really controversial in the present context. Most of those who are engaged in the disputes we are considering think that there are purely recognitional concepts of experience of the sort mentioned above — sometimes called 'phenomenal concepts' — no matter which side they occupy in the debate. (See, for example, Jackson, 1986; Block, 1995; Chalmers, 1996; Loar, 1997; Tye, 1999; Carruthers, 2000; Sturgeon, 2000.) These will be concepts that lack any conceptual connections with concepts of other kinds, whether physical, functional, or intentional.

Block and Stalnaker (1999) respond to earlier presentations of the C&J argument — as it appeared in Chalmers (1996) and Jackson (1998) — by denying the truth of premise (1). They claim that, while conditionals of the sort envisaged might sometimes be knowable from the armchair, this isn't enough to show that they are a priori. For it may be that background a posteriori assumptions of ours always play a role in our acceptance of those conditionals. While I am sympathetic to this claim (see also Laurence and Margolis, 2003), in what follows I propose to grant the truth of the first premise. In section 2 below I shall discuss some of the ways in which C&J manage to make it seem plausible.

The claim of premise (2) is that there must be an a priori conditional of the form, $(P \,\&\, T \,\&\, I) \supset M$ whenever the phenomena described in M are reductively explicable. Although I have doubts about this, too, I shall suppress them for present purposes. I propose to grant the truth of *all* of the premises, indeed. Yet there is a further suppressed assumption that has to be made before we can draw the conclusion that phenomenal consciousness isn't reductively explicable. This is an assumption about the terms in which the target of a reductive explanation must be described. And as we will see from reflection on the demands of reductive

explanation generally, this assumption is false. So it will turn out that there is no principled explanatory gap after all, and all's right with the world.

The plan of what follows is this. In section II, I discuss the sort of case that C&J are able to make in support of their first two premises, and relate their views to more traditional treatments of reductive explanation in the philosophy of science. In section III, I elaborate on the way in which purely recognitional concepts of experience generate the supposed explanatory gap. In section IV, I argue that there is a suppressed — and eminently deniable — premise that needs to be added, if we are to draw the conclusion that there is *actually* an explanatory gap. And finally, in section V, I illustrate how some recent reductive accounts of phenomenal consciousness seem to have just the right *form* to yield a complete and successful reductive explanation. (Whether any of those accounts *is* successful is of course another question.)

II Reductive Explanation and A Priori Conditionals

Chalmers (1996) makes out a powerful case in support of premise (1). On reflection it seems that we can see, just by thinking about it, that once the position and movement of every single microscopic particle is fixed (once all the microscopic facts are as they are), then there is simply *no room* for variation in the properties dealt with by macroscopic physics, chemistry, biology, and so forth — unless, that is, some of these properties are genuinely emergent, like the once-supposed *sui generis*, life-force, *élan vital*. So if we include in our description the claim that there exists nothing *except* what is entailed by the micro-physical facts, then we can see a priori that the micro-physical facts determine all the physical facts. And once we further add information about where in the micro-physically described world *I* am and when *now* is, it looks like *all* the facts (or all the facts not implicating phenomenal consciousness, at any rate) are determined. That is to say, some conditional of the form, $(P \,\&\, T \,\&\, I) \supset M$ can in principle be known to be true a priori.

Let us grant that this is so. Still, it is a further claim (made in the second premise of the C&J argument), that this has anything to do with reductive explanation. We could agree that such a priori conditionals exist, but deny that they are a requirement of successful reductive explanation. And this objection might seem initially well-motivated. For is there any reason to think that reductive explanation always aims at a suitable set of a priori conditionals? Nothing in such a claim seems to resonate with standard accounts of reductive explanation, whether those accounts are *deductive-nomological*, *ontic*, or *pragmatic* in form. So intuitively, there seems little support for the view that a priori conditionals are required

for successful reductive explanation. But actually, there is some warrant for C&J's view that the practice of reductive explanation carries a *commitment* to the existence of such a priori conditionals, at least, as will emerge when we consider existing accounts of reductive explanation.

1. The deductive-nomological account of explanation

The theory of explanation that comes closest to warranting C&J's picture is surely the classical 'deductive-nomological' model (Hempel, 1965). On this account, explanation of particular events is by subsumption under laws. An event *e* is explained once we have a statement of one or more laws of nature, *L*, together with a description of a set of initial conditions, *IC*, such that *L* and *IC* together logically entail *e*. In which case the conditional statement, '(L & IC) ⊃_ e' will be an a priori truth. When this model is extended to accommodate reductive explanation of laws, or of the properties contained in them, however, it is normally thought to require the postulation of a set of 'bridge laws,' *BL*, to effect the connection between the reducing laws *RL* and the target *T* (Nagel, 1961). The full conditional would then have the form, *(RL & IC & BL) ⊃ T*. And this, too, can be supposed to be a priori, by virtue of expressing a conceptual entailment from the antecedent to the consequent.

Notice, however, that the bridge laws will themselves contain the target terms. For example, if we are explaining the gas temperature-pressure laws by means of statistical mechanics, then one bridge principle might be, 'The mean momentum of the molecules in the gas is the temperature of the gas'. This itself contains the target concept *temperature*, whose corresponding property we are reductively explaining. There is therefore no direct support here to be had for the C&J view, that in reductive explanation there will always be an a priori conditional whose antecedent is expressed in the reducing vocabulary and whose consequent is the target being explained. For on the present model, the conditional without the bridge laws, *(RL & IC) ⊃ T*, is *not* an a priori one — there is no logical entailment from statistical mechanics to statements about temperature and pressure unless the bridge principles are included.[2]

---

2 Moreover, our reason for belief in the reducing bridge principles will be abductive, rather than a priori, of course — we come to believe that temperature (in a gas) is mean molecular momentum because assuming that it is so is simpler, and because it enables us to explain some of the processes in which temperature is known to figure.

Let me approach the same point somewhat differently. Suppose that we have achieved full understanding of what is going on at the micro-level when a gas is heated in a container of fixed volume. It should then be manifest to us that the increased momentum transmitted by the faster-moving particles to the surface of the container would have the same effect as an increase in pressure, described at the macro-level. In fact, it should be plain to us that the roles described at the micro-level — increased mean molecular momentum leading to increased transfer of momentum per unit area in a fixed volume — are isomorphic to those described at the macro-level — namely, increased temperature leading to increased pressure in a fixed volume of gas. But this isn't *yet* an explanation of the higher-level facts. Correspondence of role doesn't entail identity of role. It remains possible, in principle, that the macro-level properties might be *sui generis* and irreducible, paralleling the micro-level properties in their behavior. It is only considerations of simplicity and explanatory scope that rule this out.

But now this is, in fact, the role of the '*That's all*' clause in C&J's scheme. The micro-facts don't entail the macro-facts by themselves, C&J grant. But they will do so when conjoined with the claim that the micro-facts together with facts composed, constituted, or otherwise implied by the micro-facts are all the facts that there are.[3] What emerges, then, is that the role of the '*That's all*' clause in C&J's account is to do the same work as the bridge-principles or property identities in the framework of a classical reductive explanation, but in such a way that the target terms no longer figure on the left-hand side of the reducing conditional.

The classical deductive-nomological account of reductive explanation of properties can easily be extended to account for reductive explanation of particular facts or events, in cases where considerations of multiple realizability rule out inter-theoretic reduction or reduction of properties.

---

3 Are the implications here *conceptual* or *metaphysical*? What C&J actually say is that the '*That's all*' clause states that the world contains only the micro-facts and *what is a priori implied by* the micro-facts (317). This characterization might seem question-begging if their goal is to show that the micro-facts together with the '*That's all*' clause (and the indexicality clause) *entails* the macro-facts with a priori warrant. But it isn't. Their thought is this. We can see from the micro-facts alone that any world where such facts obtain will be a world in which there is temperature-like and pressure-like phenomena — this much is entailed a priori by the description of those facts. The micro-facts by themselves, however, don't yet rule out that in *this* world temperature and pressure are *sui generis* irreducible properties, paralleling the micro-facts in their behavior. But when we add that in this world there exists nothing *except* what is entailed by the micro-facts, then we get our required explanation — temperature and pressure are actually constituted by the micro-phenomena, because there exists nothing else to constitute them.

In place of a set of reducing laws, initial conditions, and bridge laws, we can now have reducing laws, initial conditions, and a *constituting conditional*, which states that the target phenomenon is constituted by some set of events described at the micro-level. These will together entail the presence of the target event *e*. And here, as before, C&J can claim that the constituting conditional (which contains the target terms) can be replaced by a '*That's all*' clause, yielding an a priori conditional in which the target terms figure only on the right hand side.

Before moving on, we should note that the classical account of intertheoretic reduction, as described above, soon came under pressure from those who pointed out that reduced theories often require *correction* before they can be derived from the reducing theory together with bridge principles (Sklar, 1967; Schaffner, 1976). Yet we can still regard the target properties as having been reductively explained, provided the new corrected theory is strongly analogous to the original target, and provided we can explain why the original theory works as well as it does in its domain of validity. This point will prove to be of some importance in sections IV and V, when we come to discuss the possibility of reductive explanations of phenomenal consciousness.

I conclude, then, that a priori conditionals aren't what are directly aimed at by those seeking reductive explanations within the framework of a deductive-nomological account of explanation. What is actually aimed at, are the set of reducing facts together with bridge laws, identities, or constituting conditionals that can entail the target phenomenon. But it looks like it will always be possible to construct from this an a priori conditional with the reducing facts and a '*That's all*' clause on the left-hand side, and some sort of description of the target phenomenon on the right. (This also means that the role of simplicity and other epistemic considerations has become absorbed into the left-hand side.) So C&J's claim that successful reductive explanation requires the existence of a priori conditionals would appear to be vindicated, at least within the framework of a deductive-nomological approach.

2. Ontic models of explanation

It is fair to say that a deductive-nomological approach to explanation is now a minority position. A large part of the credit for this goes to Salmon (1984, 1989), who is one of the main proponents of an opposed 'ontic' conception of explanation. On this view, to explain something isn't to offer a deductive argument for it, but rather to specify some significant part of the causal process that brought it about. And a reductive explanation of some property or process will be a description of the causal mechanism that generates that property-process.

Ontic accounts of explanation have been broadened by others to include *non*-causal relations of identity and constitution (Kim, 1974; Achinstein, 1983; Ruben, 1990). So one can explain why the pH value of some solution is changing by saying that the concentration of hydrogen ions contained in the solution is changing; and one can explain why a gas has a given temperature on the grounds that it has a given mean kinetic energy; and so forth. The relations appealed to here aren't causal ones. But the resulting account can still be described as 'ontic,' since there is no attempt to construct deductive arguments in which the explanandum figures as the conclusion. Rather, explanations proceed by telling us about the causes or the constitution of their targets.

From the perspective of ontic models it might initially seem rather unlikely that a priori conditionals will be required for successful reductive explanation. For the goal of such explanation is rather to describe the processes and mechanisms that constitute the target phenomenon. Our aim is to say something true and substantive about the world, not to construct a conditional whose truth we can see a priori. But C&J have a reply, here. For it does matter quite a lot how the target phenomena are described. Ontic explanation can't just be about relations among the properties in question *however described*. For I don't explain the rise in pH value by saying that there was a rise in pH value. It isn't identities *per se* that explain, but rather identities with a certain descriptive character.

C&J can claim, with some plausibility, that we will only ever be satisfied with a proposed reduction when the micro-phenomena mesh in the right way with the concepts used to characterize the target, in such a way as to warrant an a priori conditional. It is only when we can *see* that changing concentrations of hydrogen ions will produce just the kinds of changes distinctive of a changing pH value, that we will accept that the latter is constituted by the former. And in those circumstances it looks like a description of the micro-phenomena, combined with a '*That's all*' clause into which simplicity and other epistemic considerations have been absorbed, will a priori entail the change in pH value. And this is just what C&J claim.

By way of reinforcing this point, let us now look at the argument that C&J offer against attempts to find room for reductive explanations of phenomenal consciousness by means of bare psycho-physical identities.

3. The reductive role of identities

Block and Stalnaker (1999) argue that general considerations adduced in support of physicalism, together with correlational data discovered by neuro-psychologists, might be sufficient to warrant an *identity* between neurological facts, on the one hand, and the facts of phenomenal con-

sciousness, on the other. This would then be sufficient for phenomenal consciousness to count as reductively explained, although (a) there is no a priori conditional consisting of just micro-phenomena and a '*That's all*' clause on the left and the facts of phenomenal consciousness on the right; and (b) there is no answer to the question *why* the facts of phenomenal consciousness are constituted as they are. For as Block and Stalnaker point out, although identities are *used* in explanations, they don't, themselves, characteristically *admit of* explanation. One cannot ask, 'Why is water H_2O?' for example (note: this is not to be confused with the question, 'Why do we *believe* that water is H_2O?' which isn't problematic) — the only answer will be, 'Because that's what water *is*.'

While conceding this last point, C&J argue first, that not all identities are explanatory; and second, that they only *are* explanatory when there exists a suitable a priori conditional in which all occurrences of the target terms figure on the right-hand side. For otherwise the identity will be left as a *brute*, epistemically basic, postulate, and the higher-level property or phenomenon won't have been reductively *explained*. And they are surely right about this. Would we think that the nature of water had been explained, for example, if *all* we had to go on was the bare identity, 'Water is H_2O,' and if we couldn't use the fact of water's identity with H_2O to generate explanations of its liquidity, potability, boiling point, properties as a solvent, and so forth? And given that we *can* use the properties of H_2O to generate such explanations, we can construct an a priori conditional with the behavior of H_2O described in detail on the left (together with a '*That's all*' clause) and the claim that there exists water, on the right.

Similarly, then, in respect of Block and Stalnaker's sort of psycho-physical identity: the most that identities warranted by correlational data could explain would be the *time-course* of our phenomenally conscious experiences. But this isn't what puzzles us. We want to know what it is about such experiences that makes them available to introspective recognition, why they seem to have a distinctively subjective aspect, why they seem to their possessors to be intrinsic, ineffable, and private; and so on. Since none of this would be explained, we shouldn't count a psycho-physical identity — even if true — as a reductive explanation of phenomenal consciousness. The real explanatory work would still remain to be done. And if a brute psycho-physical identity were the best that we could hope for, then it would be reasonable to conclude that there is, indeed, an unbridgeable explanatory gap between physical facts and the facts of phenomenal consciousness.

4. Pragmatic accounts of explanation

A number of writers have claimed that explanation is a pragmatic matter, and that what makes an explanation successful is a function of the needs, knowledge and expectations of those people to whom the explanation is offered (van Fraassen, 1980; Achinstein, 1983; Lewis, 1986). Such claims come in various different strengths, and writers differ in how they think the pragmatic character of explanation relates to the accounts of explanation offered by deductive-nomological and ontic theories, of the sort discussed above. It should be plain, however, that there is nothing here that must necessarily undermine C&J's claim that successful reductive explanation requires the existence of an a priori conditional linking the reducing facts to the target. What everyone sympathetic to pragmatic accounts would insist on, however, is that whether or not an a priori conditional provides a successful reductive explanation of a target will depend crucially on the questions that puzzle us, and on whether the proffered conditional addresses those questions. This consequence is now widely accepted. And it seems to be reinforced by our discussion of the role of identities in explanation in section 2.3 above.

III Recognitional Concepts and the Explanatory Gap

It may be that we are committed to the truth of an a priori conditional, then, of the form, $(P \,\&\, T \,\&\, I) \supset M$, whenever we claim that the phenomena described in *M* are reductively explicable, or when we claim that those phenomena have been reductively explained. And for present purposes I shall accept that this is so. There exists a plausible hypothesis — endorsed by C&J — concerning the nature of our concepts for macro-phenomena which explains why such conditionals are always available (given that a reductive explanation is available). This is that such concepts are all of them broadly *functional* or *causal role* ones. We can then see that, if the micro-phenomena behave in a certain way, those roles will get filled; and we can therefore see a priori that if the micro-phenomena *are* that way, and there is nothing else, then the macro-properties must be present as well.

For example, connected with our concept of (high) temperature will be such facts as *causing pressure to rise*, *causing damage to skin*, *causing plants to wilt*, and so on. When we understand the micro-story in terms of mean molecular momentum, we can see that when the mean momentum in a gas or liquid is high there will be an increase in pressure, there will be increased damage to fragile cell-walls brought into contact with the fluid, and there will be increased evaporation from plants, causing

them to wilt. Given the details of the micro-account, we can see a priori that if there is high mean molecular momentum (and there is nothing else) then there is high temperature.

Note that to say that our concepts for macro-phenomena are broadly functional ones is *not* necessarily to say that they must be definable or analyzable into functional terms. C&J are insistent that their account need not commit them to the existence of analyses for terms like 'temperature' and 'living thing'. It may be that most such concepts don't admit of analysis at all. And yet when we deploy those concepts we can discern certain connections with other concepts a priori. C&J offer the concept *knowledge* as an example to make the point. After many decades of failed attempts to analyze the concept of knowledge, it might be reasonable to conclude that there is no such analysis to be had. But for all that, when first presented with a Gettier example, we can still see a priori that it is a case in which someone lacks knowledge. C&J's point is that our intuitions about the application-conditions of our concepts in particular cases are prior to, and more basic than, any purported general analysis (assuming that the latter is possible at all).

It is now easy to see why there can't be any a priori conditionals of the form, $(P \ \& \ T \ \& \ I) \supset C$ (at least supposing that the descriptions of phenomenal properties in C take a certain canonical form). For if some of the concepts in C are purely recognitional ones, then they will *not* be broadly functional ones. And there will then be nothing with which our micro-account can mesh conceptually to yield an a priori conditional. If our recognitional concepts for some of the qualities of our phenomenally conscious states are *purely* recognitional, then they won't carry any commitments about the circumstances in which those properties would or wouldn't be tokened, besides their phenomenal content. So when we entertain some supposed reductive story in terms of neurological events, causal roles, or intentional contents, there will be nothing to force us to conclude that in such circumstances phenomenal consciousness must be present too.

It shouldn't be claimed that *all* of our concepts for phenomenally conscious states are purely recognitional ones, of course. It may be that some of our concepts in this domain are broadly functional in character, and that some contain a combination of functional and recognitional elements (Chalmers, 1996). Consider the concept *pain*, for example. It may be that our ordinary idea of pain contains such notions as, *is caused by tissue damage* and *tends to cause grimacing and nursing of the injured body-part*, as well as including a capacity to recognize pains — straight off and without inference — as and when one has them. But it will always be possible to carve out the purely recognitional component from this concept to form a distinct concept ('*this feel*'), which will then lack any conceptual connections with role-concepts. Indeed, it may be that many

of us already possess such purely recognitional concepts, alongside a set of theoretically-embedded functional role ones.

We have, then, a pair of claims and a diagnosis. The claims are these: (1) in the case of all macro-phenomena *M* not implicating phenomenal consciousness, there is an a priori conditional of the form, $(P \,\&\, T \,\&\, I) \supset M$, and this conditional is a requirement for there to be a successful reductive explanation of the phenomena in *M*. (2) In the case of all phenomenally conscious facts and properties *C* (described using purely recognitional concepts of experience) there isn't any a priori conditional of the form, $(P \,\&\, T \,\&\, I) \supset C$ to be had; and so phenomenal consciousness doesn't admit of reductive explanation. And the diagnosis is that this difference derives from a difference in the concepts that we employ in the two domains — broadly functional, in the case of macro-phenomena, and purely recognitional in the case of phenomenal consciousness.

IV Transformed Targets and Thickly Individuated Properties

The suppressed premise in the argument for an explanatory gap, however, is that successful reductive explanations must respect the terms in which explanatory problems are posed. Our explanatory problem is, 'How can a physical system possess *this* sort of state?' where the '*this*' deploys a recognitional concept of some aspect of phenomenal consciousness. And I grant that there can be no a priori reducing conditional that has the statement, 'The system possesses *this* sort of state' on its right-hand side. Hence the appearance of an 'explanatory gap.' But what isn't yet ruled out is that we might construct an a priori conditional that has descriptions of phenomenal consciousness of some *other* sort on its right-hand side. This idea will be explored in a general way in the present section, and then illustrated with reference to recent reductive accounts of phenomenal consciousness in section V following.

Notice that in science generally, the targets of explanation don't always remain intact through the process of inquiry. In some cases we explain by *explaining away*. In the beginning our targets may be expressed in one way. But we may come to realize that they contain a false presupposition, or that the concepts with which they are expressed are in some way confused or in need of reform. For example, in astronomy we began with the explanatory problem, 'Why do the stars and the sun move across the sky in the way that they do?' But the explanation we ended up with didn't answer this question as posed. So we don't now have an a priori conditional with such-and-such facts described on the left, and the statement, 'The stars and sun move across the sky in such-and-such a way' on the right. Rather what we have is an account

of the rotation of the earth, and of the movements of the earth, sun, planets and stars in relation to one another, in terms of which we can explain why the sun and stars *appear* to move across the sky in the way that they do.

For another example of the same general type, consider evolutionary biology. Here we began (pre-Darwin) with an explanatory question: why do species exist? But now (post-Darwin) we see that there are no such things as species in the original intended sense. Rather, there exist a great many populations of individual organisms spread out over space and time, that resemble one another more or less closely, and that stand in various inheritance relations to one another. The idea of *species* as some sort of underlying unifying essence has now been dropped. And what gets explained instead are the ways in which similarity relations amongst individuals shift over time, given facts of inheritance and facts about survival and reproduction. So here, too, there is no a priori conditional available to us that has a body of explanatory theory on the left-hand side and the statement (as originally intended), 'There are different species of living thing' on the right.

What makes these examples work, is that in the course of inquiry, and in the course of adopting our explanatory theories, we have realized that our initial questions made false presuppositions. So we have shifted to a new set of questions to which we can now provide direct answers. And it might be objected against any attempt to model our failures to produce a reductive explanation of phenomenal consciousness on this, that in the above cases we *do* finish with a priori conditionals with everything that we *currently* believe to be true on the right-hand sides. In the case of phenomenal consciousness, in contrast, the proposal would presumably have to be that all of our beliefs involving purely recognitional concepts would need to be left outside the scope of the explanatory conditional.

This is a fair point, and a significant difference. But in reply we can claim that the moral of the examples is really this: explanations succeed when there is nothing left to explain. Explanations are complete when every question that we *want* answered has *been* answered. And reflection can make us see that there are some questions, that we might initially have been inclined to ask, that no longer require answers. (The question, 'Why am I lucky?' might be an example falling into this category.) And this is what many proposed reductive explanations suggest in respect of phenomenal consciousness, as we shall see in the next section. They offer a reductive account from which we could construct an a priori conditional concerning many facts about phenomenal consciousness. And at the same time they offer a reductive account of why there *can't* be any such reducing conditionals with statements containing purely recognitional concepts on their right-hand sides. So we are supposed to see, in

the end, that every question that requires an answer has received an answer.

C&J might reply that even if — pragmatically — explanation stops when all the questions we want answered have been answered, it is a further constraint on the success of a reductive explanation that *every* fact at the target level *could* be reductively explained *in principle*. In which case, by conceding that there are facts of phenomenal consciousness expressed using recognitional concepts that can't be reductively explained, we have accepted that phenomenal consciousness itself can't be explained.

Instead of challenging the premise of this argument, let me just accept it, and go on to draw some distinctions. First, as we noted in section II.2, whether or not an explanation is successful can turn crucially on the way that the target is described, even from the standpoint of an ontic account of explanation. Some of the descriptions that figure on the right-hand side need to be drawn from the same *level* as the target, at least — e.g. involving phenomena that in one way or another pertain to *temperature*, described as such. *Which* descriptions from a given level are the relevant ones, though? Surely not *all*. The requirement cannot be that a successful reductive explanation should be able to generate *all* descriptions of the target phenomenon; for there will be infinitely many (indeed, uncountably many) of these. So even idealizing for limitations on time, memory and so on (C&J, 334), reductive explanation would become impossible. The relevant descriptions are the ones that puzzle us, perhaps, or that seem central to the characterization of the phenomena in question.

Now let me introduce a distinction between facts that are *thickly* and *thinly* individuated. In the first — *thick* — sense of 'fact,' one fact may be the subject of many distinct thoughts. Here facts might be thought of as ordered n-tuples of individuals, properties, and relations. But in the second — *thin* — sense, facts are partly individuated in terms of the thoughts used to express them. In which case, whenever we use distinct concepts in characterizing a fact, we have thereby described a *distinct fact*. So in the thick sense, the fact that I am holding up five fingers, and the fact that the number of fingers I am holding up is the smallest prime number larger than three, are the *same* fact. But in the thin sense, these are two distinct facts. Notice that we can draw the thick-thin distinction, not just *across* levels (with one given underlying thickly individuated fact being picked out by two distinct descriptions at a higher level, or by descriptions at two different levels), but also within the *same* level. For example, it is one and the same thickly individuated fact that is picked out by, 'John is thinking about ex-President Nixon' and, 'John is thinking about the President who resigned over the Watergate affair'.

Given the distinction between thick and thin facts, we can claim this. While it is a constraint on reductive explanation that the target should

be described at the appropriate 'level'; and while it is a constraint on complete success in explanation that every *thickly individuated* fact at the target level should receive an explanation; it cannot be a rational constraint on explanation that every *thinly individuated* fact should be explained. There are just too many of them (infinitely many), for one thing. The suppressed assumption in the C&J argument for an explanatory gap can now be expressed more precisely. It is the assumption that reductive explanations must require a priori conditionals *in which all the thinly individuated facts concerning the target phenomena figure on the right-hand side*.

What I claim, then, is that this suppressed premise is false. A reductive explanation of phenomenal consciousness could be successful by doing the following. It could explain all that *needs* explaining at the target level, leaving no significant question unanswered; and it could be extended (in principle) to explain every thickly individuated fact in the target domain, described at the target level. But there will remain some *thinly individuated* facts (viz., those expressed using purely recognitional concepts) that remain outside the scope of the resulting a priori conditional. Yet our reductive account can at the same time explain just *why* such statements must remain outside the scope of the conditional. This, I argue, would be complete success.

V The Form of Reductive Explanations of the Phenomenal

Phenomenally conscious properties can be characterized purely recognitionally, from a first-person perspective; which makes it hard to see initially how any reductive story could connect appropriately with those properties. But it is important to realize that phenomenally conscious properties *also* admit of third-personal characterization. (The idea I make use of here is a derivative of Dennett's 1991 notion of *hetero-phenomenology*.) Most obviously, we can say that these are properties that are available to introspective recognition. We can say, too, that these properties have a 'fineness of grain' that gives them a richness well beyond our powers of description and categorization. And we can add that people are strongly inclined to think of phenomenally conscious states as possessing intrinsic — that is, non-relational and non-intentional — properties, that are available for introspective classification; that people are inclined to think of these properties as ineffable and private; and that we are inclined to think that we have incorrigible, or at least privileged, knowledge of them.

Bundling these third-person characterizations into a third-person concept of phenomenal consciousness, we can then pick out each thickly

individuated fact designated through the application of a purely recognitional concept by saying, 'It is the phenomenally conscious state that he/she is introspectively recognizing right now'. The claim is that each such fact — together with the various puzzling properties that make up the third-person concept of phenomenal consciousness — can in principle receive a reductive explanation.

Such third-person characterizations seem tailor-made for explanation from the perspective of 'intentionalist' or 'representationalist' theories of phenomenal consciousness, indeed — whether of a first-order sort (Kirk, 1994; Dretske, 1995; Tye, 1995, 2000) or of a higher-order kind (Dennett, 1978, 1991; Lycan, 1987, 1996; Carruthers, 2000). This is not the place to develop this claim in any detail; and of course there are important differences between the different reductive accounts on offer here.[4] But notice, for example, that an appeal to the 'analog' or 'non-conceptual' intentional content of our perceptual states can explain the fineness of grain associated with phenomenal consciousness. And notice, too, that any property that is the object of a bare-recognitional concept will be apt to seem intrinsic to someone deploying that concept.

Intentionalist explanations aren't yet *micro*-explanations, of course. So it is presupposed, first, that the facts of intentional psychology will in turn admit of reductive explanation in physical terms; and second, that intentional psychology can be carved off from anything involving phenomenal consciousness. Both presuppositions are to some degree controversial. There is much debate about whether, and if so how, intentional content can be reductively explained (Millikan, 1984; Fodor, 1990; Searle, 1992, 1997). And some deny that intentional content can be understood and characterized apart from phenomenal consciousness (Searle, 1992, 1997). But I don't need to enter into disputes about the naturalization of intentional content here. For my goal is not to defend the view that phenomenal consciousness can actually be reductively explained by micro-physics, but just that it is reductively explicable *in principle*.

However, I do need to claim that Searle is mistaken in thinking that intentional content itself presupposes phenomenal consciousness. For otherwise the suggestion that the puzzling features of phenomenal

---

4 One important dimension of difference concerns the question of how much of our characterization of phenomenal consciousness — e.g. its rich and fine-grained character, or its possession of intrinsic non-relational properties (qualia) — is *explained*, and how much is explained *away* as resulting from some sort of illusion. (Many have alleged that Dennett's 1991 should really have been entitled, *Consciousness Explained Away*, for example.)

consciousness can (even in principle) be explained by appeal to intentional content will be spurious. I shan't argue for this here, however, since Searle's position is endorsed by hardly anyone else working in the field (appeals to non-conscious intentional states are now routine in both philosophy and cognitive science), and since it isn't endorsed by C&J in particular.

It may be objected that intentionalist explanations don't in any case touch the core, or the defining feature, of phenomenal consciousness. This is its 'what it is likeness' (Nagel, 1974), that can only be characterized using our first-person recognitional concepts themselves. Yet we can have good reason to think, surely, that the properties picked out by our first-person recognitional concepts are the very same properties as those that figure in the third-person characterizations sketched above. And then a reductive account of those properties will be what we seek, provided it can answer all the questions that puzzle us. A successful reductive explanation that focuses on the third-person characterizations can give us good reason to think that phenomenal consciousness *per se* has been reductively explained.

Most importantly, a representationalist approach can deliver a third-person account of our recognitional concepts for the target properties that explains why, to anyone employing those concepts, the explanatory gap will seem unbridgeable. (For detailed proposals of this sort see Tye, 1999, and Carruthers, 2000.) For example, if we possess *purely* recognitional concepts of the form, '*This* type of experience' — with no a priori links to functional-role concepts or intentional concepts, in particular — then no matter what reductive account we are offered in causal-role or intentional terms, we will still be able to think, 'All that might be true without *this* type of experience being present.' But the property picked out by 'This experience' might be, nevertheless, the very same as the one given in the causal-intentional theory. And the success of that theory in accounting for the various third-person characterizations of the puzzling features of phenomenal consciousness can give us good reason to believe that it is.

The form that these various reductive proposals take, then, is this. There is a micro-story (in this case cast in causal-intentional terms) from which we can see a priori that in any world in which it is true, the various puzzling facts about phenomenal consciousness will be true, in so far as those facts can be described without using our purely recognitional concepts. That is, we can see a priori that in any such world, people will be able to know immediately of the presence of their experiential states through introspective recognition, and they will be inclined to think that their experiential states possess properties that are ineffable, intrinsic, and private. Moreover, we can see a priori that in such a world, people will possess concepts for their experiences that (by virtue of their purely

recognitional character) will rule out any sort of a priori reducing conditional that has statements expressed using those concepts on the right-hand side.

Does it follow from the micro-story that in our world phenomenal consciousness is constituted by the truth of that story? No, not yet — any more than it follows from the micro-story alone that in our world temperature in a gas is constituted by mean molecular momentum. Here (as in the case of temperature) we need to add a '*That's all*' clause, warranted in the same way by considerations of simplicity and explanatory power. And then we can construct an a priori conditional of the form, $(P \ \& \ T \ \& \ I) \supset C$, where what figures in C aren't statements employing our recognitional concepts of experience, but rather third-person descriptions of the various puzzling facts about phenomenal consciousness (including, note, the fact that I can still think the thought, 'All of $P \ \& \ T \ \& \ I$ might be true, while I nevertheless lacked *this*,' where 'this' expresses a recognitional concept of experience). And a third-person description of every phenomenally conscious property that is the object of such introspective recognition can also figure in C.

Notice that proposed reductive explanations of this form will only work by re-configuring the terms in which the target is expressed. Instead of asking, 'How can any physical system have *this* sort of experience?' (deploying a recognitional concept in the explanandum), we now ask, 'How can any physical system have states that seem ineffable and private, etc., and which persistently induce the feeling of an explanatory gap?' But it is not at all unusual for successful explanations to require that their targets be reconfigured in this sort way. In astronomy, as we saw earlier, we had to stop asking, 'Why do the sun and the stars move across the sky in the way that we do?' and ask instead, 'Why do the sun and stars *appear* to move as they do?' The temptation to *see* the sun as moving still persists. But we no longer take it seriously. For we know that a rotating earth, together with a visual system that takes the earth as its frame of reference in perceptions of motion, will produce just such an impression.

Reductive explanations are successful when (a) all of the questions that puzzle us are answered, either directly, or indirectly by showing why the facts should *seem* a certain puzzling way to us when they are not; and when (b) every thickly individuated fact described at the target level can be reductively explained. And this is just what is claimed by the various causal-intentional reductive theories of phenomenal consciousness on the market. Where the C&J argument goes wrong, is in its assumption that reductive explanations require a priori conditionals in which the target phenomena *as initially characterized* are described on the right-hand sides, and in which *all* the *thinly individuated* target facts figure on the right-hand sides.

VI Conclusion

For purposes of argument I have conceded to C&J that successful reductive explanations require a priori reducing conditionals containing references to the target properties on their right-hand sides. But I have insisted that reductive explanation can often require a re-working of the terms in which the target phenomena are conceptualized, or the terms in which our explanatory questions are posed. And I have insisted, too, that while all the target facts (thickly individuated, but described at the appropriate level) need to figure on the right-hand sides of such conditionals, it *isn't* true that all *descriptions* of such facts need to be capable of figuring there. When these points are brought into proper focus, it is plain that there is no obstacle of principle, here, to the provision of a reductive explanation of phenomenal consciousness. Whether such an explanation can in fact be provided is, of course, a topic for another occasion.[5,6]

Received: February 2003

References

Achinstein, P. 1983. *The Nature of Explanation*. New York: Oxford University Press.

Balog, K. 1999. 'Conceivability, Possibility, and the Mind-Body Problem,' *The Philosophical Review* **108** (1999) 497-528.

Block, N. 1995. 'A Confusion about the Function of Consciousness,' *Behavioral and Brain Sciences* **18** (1995) 227-47.

Block, N. and Stalnaker, R. 1999. 'Conceptual Analysis, Dualism and the Explanatory Gap,' *The Philosophical Review* **108** (1999) 1-46.

Carruthers, P. 2000. *Phenomenal Consciousness*. New York: Cambridge University Press.

Chalmers, D. 1996. *The Conscious Mind*. New York: Oxford University Press.

---

5 See Carruthers (2000) for detailed discussion, and also for development and defense of a novel form of dispositionalist higher-order thought theory, comparing it with alternatives.

6 Thanks to George Botterill, Jeffrey Bub, David Chalmers, Lindley Darden, Mathias Frisch, Frank Jackson, Rosanna Keefe, Stephen Laurence and Paul Pietroski for comments on, and/or insightful discussion of, earlier drafts of this paper. I am especially grateful to an anonymous referee for criticisms that occasioned a complete rebuilding of the paper, 'from-the-ground-up'.

Chalmers, D. 1999. 'Materialism and the Metaphysics of Modality,' *Philosophy and Phenomenological Research* **59** (1999) 473-96.

Chalmers, D. and Jackson, F. 2001. 'Conceptual Analysis and Reductive Explanation,' *The Philosophical Review* **110** (2001) 315-60.

Dennett, D. 1978. 'Toward a Cognitive Theory of Consciousness,' in *Minnesota Studies in the Philosophy of Science* **9**, C. Savage, ed.

Dennett, D. 1991. *Consciousness Explained*. London: Allen Lane.

Dretske, F. 1995. *Naturalizing the Mind*. Cambridge, MA: The MIT Press.

Flanagan, O. 1992. *Consciousness Reconsidered*. Cambridge, MA: MIT Press.

Fodor, J. 1990. *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. 1998. 'There are no Recognitional Concepts, not even RED,' in his *In Critical Condition*. Cambridge, MA: The MIT Press.

Hempel, C. 1965. *Aspects of Scientific Explanation*. New York: Free Press.

Jackson, F. 1982. 'Epiphenomenal Qualia,' *Philosophical Quarterly* **32** (1982) 127-36.

Jackson, F. 1986. 'What Mary Didn't Know,' *Journal of Philosophy* **83** (1986) 291-5.

Jackson, F. 1998. *From Metaphysics to Ethics*. New York: Oxford University Press.

Kim, J. 1974. 'Non-Causal Connections,' *Nous* **8** (1974) 41-52.

Kirk, R. 1994. *Raw Feels*. New York: Oxford University Press.

Laurence, S. and Margolis, E. 2003. 'Concepts and Conceptual Analysis,' *Philosophy and Phenomenological Research* **67** (2003).

Levine, J. 1983. 'Materialism and Qualia: The Explanatory Gap,' *Pacific Philosophical Quarterly* **64** (1983) 354-61.

Levine, J. 1993. 'On Leaving Out What It's Like,' in *Consciousness*, M. Davies and G. Humphrey, eds. New York: Blackwell.

Levine, J. 2001. *Purple Haze*. New York: Oxford University Press.

Lewis, D. 1986. 'Causal Explanation,' in his *Philosophical Papers*, vol. 2. New York: Oxford University Press.

Loar, B. 1997. 'Phenomenal States,' in *The Nature of Consciousness*, N. Block, O. Flanagan and G. Güzeldere, eds. Cambridge, MA: MIT Press.

Lycan, W. 1987. *Consciousness*. Cambridge, MA: The MIT Press.

Lycan, W. 1996. *Consciousness and Experience*. Cambridge, MA: The MIT Press.

McGinn, C. 1991. *The Problem of Consciousness*. New York: Blackwell.

Millikan, R. 1984. *Language, Thought, and Other Biological Categories*. Cambridge, MA: MIT Press.

Nagel, E. 1961. *The Structure of Science*. New York: Routledge.

Nagel, T. 1974. 'What is it Like to be a Bat?' *Philosophical Review* **83** (1974) 435-50.

Ruben, D.-H. 1990. *Explaining Explanation*. New York: Routledge.

Salmon, W. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton, NJ: Princeton University Press.

Salmon, W. 1989. 'Four Decades of Scientific Explanation,' in *Minnesota Studies in Philosophy of Science* **13**, P. Kitcher and W. Salmon, eds.

Schaffner, K. 1976. 'Reductionism in Biology,' in *Philosophy of Science Association 1974*, R. Cohen, et al., eds. Dordrecht, Netherlands: Reidel.

Searle, J. 1992. *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Searle, J. 1997. *The Mystery of Consciousness*. A New York Review Book.

Sklar, L. 1967. 'Types of Inter-Theoretic Reduction,' *British Journal for the Philosophy of Science* **18** (1967) 109-24.

Sturgeon, S. 1994. 'The Epistemic View of Subjectivity,' *Journal of Philosophy* **91** (1994) 221-35.

Sturgeon, S. 2000. *Matters of Mind*. New York: Routledge.

Tye, M. 1995. *Ten Problems of Consciousness*. Cambridge, MA: MIT Press.

Tye, M. 1999. 'Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion,' *Mind* **108** (1999) 705-25.

Tye, M. 2000. *Consciousness, Color and Content*. Cambridge, MA: The MIT Press.

van Fraassen, B. 1980. *The Scientific Image*. New York: Oxford University Press.

Wittgenstein, L. 1953. *Philosophical Investigations*. New York: Blackwell.