Paul Livingston

Consciousness, Structuralism and Open Questions

In this paper, I want to argue for two claims. The first is that arguments for the irreducibility of consciousness can be put in the form of *open question arguments*: that is, arguments of a certain sort which turn upon the openness, or the intelligibility, of certain explanatory questions under the assumption that certain sorts of putatively explanatory facts are given. Second, I'll claim that putting the irreducibility arguments in this form has the benefit of demonstrating the actual form of the conclusion they establish, and thus help to distinguish this actual conclusion from other, less far-ranging conclusions that have been drawn from them. Ultimately, I hope to show that the irreducibility arguments, if understood in this way, can be construed as establishing conclusions even stronger than those that they are usually taken to: for they succeed in establishing, not only that consciousness cannot be explained in terms of *physical* matter and forces, but that there is and can be no *naturalist* (or natural-scientific) explanation of consciousness at all. For, as the irreducibility arguments show if we put them in the form of open question arguments, consciousness is not anything like a "natural" property or feature (either of physical or phenomenal "stuff"), but rather a (logically) different *kind* of thing altogether.

The first and best-known "open question" argument (and the one usually known by that name) is the one given by G.E. Moore in *Principia Ethica* (1903):

> The hypothesis that disagreement about the meaning of good is disagreement with regard to the correct analysis of a given whole, may be most plainly seen to be incorrect by consideration of the fact that, whatever definition may be offered, it may always be asked, with significance, of the complex so defined, whether it is itself good. To take, for instance, one of the more plausible, because one of the more complicated of such proposed definitions, it may easily be thought, at first sight, that to be good may mean to be that which we desire to desire. Thus if we apply this definition to a particular instance and say 'When we think that A is good, we are thinking that A is one of the things which we desire to desire,' our proposition may seem quite plausible. But, if we carry the investigation further, and ask ourselves 'Is it good to desire to desire A?' it is apparent, on a little reflection, that this question is itself as intelligible, as the original question, 'Is A good?' – that we are, in fact, now asking for exactly the same information about the desire to desire A, for which we formerly asked with regard to A itself. But it is also apparent that the meaning of this second question cannot be correctly analysed into 'Is the desire to desire A one of the things which we desire to desire?': we have not before our minds anything so complicated as the question 'Do we desire to desire to desire to desire A?' Moreover any one can easily convince himself by inspection that the predicate of this proposition – 'good' – is positively different from the notion of 'desiring to desire' which enters into its subject: 'That we should desire to desire A is good' is *not* merely equivalent to 'That A should be good is good.'

The argument may be reconstructed as follows, establishing, as Moore says, that there is no analysis (or definition) of "good" in terms of any "natural property" or set thereof:

1. Suppose there were an analysis of "good" as some natural property, A (e.g. being what we desire to desire).
2. Then, given an instance, a of A (e.g., something that we desire to desire), it would not be an open question to ask whether a is good.
3. But given any such property A and any instance a, it is still an open question, even given that a is A, whether a is good.
4. Therefore there is no natural property A that analyses "good" (i.e., "good" is undefinable).

As Moore says, the argument aims to establish the untenability of understanding "good" as analyzable in terms of any natural property or feature, or combination thereof, and so to support the diagnosis of what Moore calls the "naturalistic fallacy," the fallacy of understanding "good" to designate some *natural* property.

Although Moore's argument is usually discussed as "the" open question argument, there is another historically important argument of a similar form that bears comparison and that also may be understood as turning, in a similar way, on the openness of a certain kind of question, given any natural facts whatsoever. This is Frege's argument, in "The Thought," for the undefinability of truth. The argument begins by rejecting any definition of truth as "correspondence" in a certain respect and finishes by rejecting the claim that truth can be defined at all:

> But could we not maintain that there is truth when there is correspondence in a certain respect? But which respect? For in that case what ought we to do so as to decide whether something is true? We should have to inquire whether it is *true* that an idea and a reality, say, correspond in the specified respect. And then we should be confronted by a question of the same kind, and the game could begin again. So the attempted explanation of truth as correspondence breaks down. For in a definition certain characteristics would have to be specified. And in application to any particular case the question would always arise whether it were *true* that the characteristics were present. So we should be going round in a circle. So it seems likely that the content of the word 'true' is *sui generis* and indefinable.[1]

Though it concerns truth and not the good, Frege's argument can be put in a similar form to Moore's:

1. Suppose there were an analysis of "true" as some natural property or relation, B (e.g. a relation of correspondence in some naturally definable respect)

[1] "Thought", p. 327.

2. Then, given an instance, b of B (e.g., a sentence or thought that resembles its object in some naturally definable respect), it would not be an open question to ask whether b is true.
3. But given any such property or relation B and any instance b, it is still an open question, even given that b has property B (or bears relation B to something else), whether b is true (or: whether it is true that b has the property B)[2]
4. Therefore there is no natural property A that analyses "true" (i.e., "true" is undefinable).

Here, as in Moore's argument, the fact that the question about any specific instance of a (natural) property or relation remains open for any (natural) property or relation thought to define the target notion shows that there is *no* definition of the target notion in terms of any (natural) property or relation.

There is another kind of argument, of a more recent vintage, that can also naturally (as has been suggested) be put in the form of an open question argument. The more recent kind of argument concerns consciousness, and is usually used to establish an anti-physicalist conclusion:

1. Suppose a state's being conscious were identical with its having some physical property or being of some physical state-type, C
2. Then, given an instance c of C (e.g. a brain state having property C), it would not be an open question whether c is a conscious state (or whether an organism in state c is in a certain conscious state)
3. But given any such property or state-type C and state c, it is still an open question, even given that c is C, whether c is a conscious state (or whether an organism in state c is in a certain conscious state)
4. Therefore there is no physical property or state-type C that is identical with consciousness

At least two anti-physicalist arguments that, it appears, might be put in roughly this form are Chalmers' "zombie" argument for the nonsupervenience of consciousness on physical facts and Jackson's "knowledge" argument. For the "zombie" argument, we imagine zombies that are identical to us in all physical and functional respects, and note that there is still an open question whether the zombies have phenomenal consciousness (i.e. that it is conceivable that they do not). For Jackson's argument, we imagine Mary knowing all the physical facts, and we note that there is still an open question for her as to

[2] There is a slight wrinkle here in that Frege specifies the open question, given a relation B that putatively defines "truth" and two individuals, P and Q, standing in relation B, not simply as the question whether P is true (which would be closer to Moore's version of the open question) but rather as the question *whether it is true* that P bears B to Q. But the point is much the same: the openness of this further question shows that given the purported definition there is still an open question with respect to the very notion that is supposed to be defined, which there should not be if the definition is successful.

what it is like to be in a certain phenomenal state (i.e. whether any of the physical state-types with which she is already familiar is identical with that phenomenal state type).

It may be that other anti-physicalist arguments can also be put into this form; at any rate, these two clearly turn on the existence of an open question. In "Consciousness and it Place in Nature," Chalmers suggests that the "zombie" argument and Jackson's argument can be seen as having a common form that is close to what I've presented above (what he calls the form of an "epistemic" argument against materialism) and in "Facing up to the Problem of Consciousness" he suggests explicitly that the underlying problem might be put as one of an open question:

> When it comes to conscious experience, this sort of explanation [viz., explanation in terms of "the performance of functions"] fails. What makes the hard problem hard and almost unique is that it goes beyond problems about the performance of functions. To see this, note that even when we have explained the performance of all the cognitive and behavioral functions in the vicinity of experience – perceptual discrimination, categorization, internal access, verbal report – a further unanswered question may remain: *why is the performance of these functions accompanied by experience*?

Similarly, Joseph Levine has suggested in various places that the argument for the familiar claim that there is an "explanatory gap" between physical facts and consciousness may well be put as an open question argument: given all of the physical facts, there remains an open question about *why* a certain physical state is correlated with consciousness or with a conscious state of a certain type, even if it is in fact the case that it is correlated.[3] [4]

For all of these anti-physicalist arguments, the idea that there is an open question even given any possible physical explanation plausibly captures the intuition that it is at the root of the argument. Thus, the general open question argument I have presented above may be thought to capture well what is common to all these anti-physicalist arguments. Beyond this commonality, moreover, there are certain advantages to putting each of these arguments in the form of an open question argument. One

[3] See, e.g., Levine: "Phenomenal Concepts and the Materialist Constraint." Michael Pelczar discusses the connection between the Jackson's argument and Moore's argument in "The Knowledge Argument, the Open Question Argument, and the Moral Problem," arguing that both arguments fail for similar reasons. Uriah Kriegel has also noted this connection between arguments for the irreducibility of consciousness and open question arguments.

[4] In The *Conscious Mind* (pp. 83-84), Chalmers suggests that there is a disanalogy between the conclusion that Moore's argument is meant to establish and the conclusion of the nonsupervenience argument in that, whereas a zombie world is conceivable, it is not possible to conceive of a world identical to ours in all non-moral respects but in which moral facts are different. However, if the current analysis is right, the presence of an open question argument in both cases points to a similarity that cuts deeper than this apparent disanalogy about the conceivability of various worlds. This is because the open question argument itself motivates, in both cases, the claim that there is something about the relevant phenomena that is "over and above" the natural facts, and thus (if the argument is successful) there is no need for a further conceivability intuition to motivate this claim.

advantage is that the open question argument does not appear to depend in any direct way on an "inference" from epistemic conceivability to metaphysical possibility, as (other) forms of the "epistemic argument" do. In each case, it is true that the argument does depend in a certain sense on the claim that certain scenarios – in which certain physical states are present but certain phenomenal states are not – are indeed conceivable. But on the present approach, to say that these scenarios are conceivable is just to restate that there is an open question about the presence of the phenomenal states even given all of the physical states, and *this* claim is already (at least in part) an "ontological" one. That is, to agree that there is an open question about phenomenal experience, even given all the physical facts, is already to recognize that the character of phenomenal experience is something "above and beyond" the physical facts. There is no special further inference needed from epistemic conclusions (e.g. about conceivability) to metaphysical ones about ontology. Similarly, and for related reasons, the open question argument does not depend on any contentious metaphysical framework (for instance any particular conception of possible worlds or supervenience). Rather, it is formulated as a straightforward argument about explanatory success, tending to show that phenomenal consciousness *resists* explanation of a certain sort, and that the reason for this resistance is (*prima facie* at least) as at least as much rooted in ontological issues as it is in epistemology.

The biggest challenge faced by the open question argument, as I have put it, arises from the thought that phenomenal consciousness might in fact be identical with some physical or functional state, and that this identity might be *a posteriori*. If this were the case (and the identity between phenomenal consciousness and some physical state were similar, e.g., to the identity between water and H_2O) then (so an objector might argue) there is no inference from the fact that there is an open question to the conclusion that consciousness is not in fact identical with a physical or functional state. For the question might be "open" only because science has not yet closed it: that is, it may *appear* that an explanation of all the physical facts leaves the nature and existence of consciousness open, whereas in fact (as an a posteriori matter) it does not, as we would discover when the actual identity is in fact discovered. If this were the case, then the open question argument would establish, at best, that our understanding of consciousness is not yet complete rather than (as intended) that *no* physical or functional explanation of consciousness can succeed.

This objection trades on the phenomenon of *a posteriori* necessity, discussed by Kripke in *Naming and Necessity*, and on the further thought that the identity between consciousness and a certain physical or functional state might be similar to other cases of *a posteriori* identity. However, Kripke himself argues in the last pages of *Naming and Necessity* that the question of consciousness is in important respects *disanalogous* with other cases in which identities have been discovered *a posteriori*, and there are various ways to argue that the idea of *a posteriori* identity doesn't help the physicalist to resolve the underlying problem with respect to consciousness. One way is to argue, as Chalmers does in *The Conscious Mind*, that it is epistemic or "primary" intensions, rather than metaphysical or "secondary" intensions, that are relevant to the explanation of consciousness. This is because, as he says there (p. 69), it is the primary intension rather than the secondary one that determines what is to be explained, and so sets the criteria for successful explanation. If this is right, then issues about the success of

various types of explanation for various types of phenomena can be addressed in a largely *a priori* fashion, and the resolution of these issues is not affected by the possibility of *a posteriori* discoveries. In particular, as Chalmers suggests, it is still possible to argue that, notwithstanding the possibility of *a posteriori* discoveries, consciousness fails to logically supervene upon the physical facts.

Still, one might suspect that the open question argument, as I have presented it, moves too quickly from explanatory to ontological issues. Why should the apparent openness of a certain type of explanatory question, or what Moore called the "absence of an analysis", carry the metaphysical weight of establishing that no identification between consciousness and a physical state is possible, when other sorts of identities have been established *a posteriori* by scientific inquiry? In response, it is helpful to note some further disanalogies between the consciousness-physical state and the water-H2O case. One disanalogy has to do with the sort of explanation (or analysis) it is reasonable to expect *in advance of* empirical inquiry. Although it was not always known that water is H_2O, it was reasonable to expect even in advance of this discovery that water could be explained as some sort of phenomenon of matter, perhaps even as composed of something like compounded physical atoms. Even in advance of *the a posteriori* discovery that water is H_2O, therefore, it would not have been possible to argue that any purported physical explanation of water still leaves an open question. This is disanalogous with the consciousness case, where it is apparently (if the open question argument is successful) possible to argue precisely this. Moreover, as Levine has quite rightly pointed out, given that water *has* been discovered to be identical to H_2O, there is today *no longer* any open question taken seriously by educated adults about *why* water should be H_2O.[5] If the open question argument is right, this is again disanalogous with the consciousness case, since given any purported explanation of consciousness as identical to *any* physical phenomenon or phenomena, there will still remain the open question. Another standard way to put this is that consciousness gives rise to an "intuition of distinctness" from the physical that is not present in the case of water, or virtually any other (non-phenomenal) substance, state, or process.

What, then, is the basis for this striking difference in terms of explanation between phenomenal consciousness and other familiar processes and events? Interestingly, although Kripke is of course the philosopher most responsible for the contemporary discussion of *a posteriori* identity that is sometimes applied by physicalists to the explanation of consciousness, in the closing pages of *Naming and Necessity* he gives an argument against physicalism that may itself be taken to be a form of the open question argument I've considered above.

In particular, Kripke argues that consciousness is disanalogous with other phenomena such as water or heat in that, whereas these phenomena may be identified by means of some of their contingent properties (in particular, the contingent property of causing certain *phenomenal experiences* in us), consciousness is always picked out, in our first-person reference to it, by means of its *necessary* properties (i.e., the very phenomenal properties that constitute it). Thus, whereas it is possible, for

[5] In "Phenomenal Concepts and the Materialist Constraint"

instance, in the case of heat, to make the *a posteriori* discovery that *the very phenomenon* that is picked out as the normal *cause* of our heat-sensations (viz., by means of one of its *contingent* properties, the property of causing this sort of sensation in us) is itself identical with molecular motion, this sort of a posteriori discovery of identity is not apparently possible with respect to consciousness. This is because phenomenal states are picked out by means of their phenomenal properties, which are *necessary* properties of those states. (In particular: "…pain, unlike heat, is not only rigidly designated by 'pain' but the reference of the designator is determined by an essential property of the referent.")

Kripke's argument against physicalism is sometimes put as simply a Cartesian-style argument from the "intuition of distinctness" between mind and body, or the conceivability of a distinction between mind and body, to the real possibility of their distinction. However, there is more to the argument than this. In particular, it turns specifically on the distinction in *epistemic* possibilities that are open in the case of consciousness vs. those that obtain in the case of established *a posteriori* identities. Though these are all necessary identities (Kripke argues that there is no such thing as a "contingent identities")[6], and so it is metaphysically impossible that heat, for instance, might not have been molecular motion, it is *epistemically* possible (as Kripke willingly grants) that heat might have turned out to be something other than molecular motion. That is, if the sensations we get from heat had turned out to be caused by some other phenomenon, we would have put this situation as one in which "heat" (i.e. the referent of the term "heat") turned out to be something other than molecular motion. There could, in other words, have been an initial epistemic situation qualitatively identical to the actual one, but in which heat sensations turned out later not to be caused by molecular motion. This is the very possibility that makes the heat-molecular motion identity, though necessary, *a posteriori*, but it is just this possibility that does *not* obtain in the consciousness-brain state case. This is because, as Kripke argues, it is incoherent to suppose that there could be an epistemic situation qualitatively identical to one in which a subject has a particular sensation, but in which *that very sensation* doesn't exist:

> The trouble is that the notion of an epistemic situation qualitatively identical to one in which the observer has a sensation S simply is one in which the observer had that sensation. The same point can be made in terms of the notion of what picks out the reference of a rigid designator. In the case of the identity of heat with molecular motion the important consideration was that although 'heat' is a rigid designator, the reference of that designator was determined by an accidental property of the referent, namely the property of producing in us the sensation S. It is thus possible that a phenomenon should have been rigidly designated in the same way as a phenomenon of heat, with its reference also picked out by means of the sensation S, without that phenomenon being heat and therefore without its being molecular motion. Pain, on the

[6] It is somewhat ironic, given some of the uses to which Kripke's notion of the "necessary a posteriori" has been put by physicalists, that as he makes clear in the "Preface" to *Naming and Necessity*, one of his major polemical goals in the book as a whole was to dispute the notion of "contingent identity" that had been thought, e.g. by J.J.C. Smart in "Sensations and Brain Processes", to support a physicalist identity theory, and so to argue against the (then common) claim that mental states and brain states might be "contingently identical".

> other hand, is not picked out by one of its accidental properties; rather it is picked out by the property of being pain itself, by its immediate phenomenological quality.

There is a sense, in other words, in which phenomenal states are essentially self-designating or self-presenting, and it is this that distinguishes them from the candidates for *a posteriori* identity. In particular, because of this self-designating character of phenomenal states, there is not (in the case of phenomenal states) the gap between the (usually contingent) properties used to designate the phenomenon in cases like the case of heat and the phenomenon itself, which in *those* cases makes an a posteriori discovery of identity possible. Thus, Kripke argues, if phenomenal states *were* (necessarily) identical with some physical states, then there would not be an open question for *a posteriori* discovery about the basis for their identity. However, given any claim of identity, there is such an open question. So it appears to follow that phenomenal states are not identical with any physical states.

II

I have argued that familiar arguments for the irreducibility of consciousness or its non-physical status can usefully be put in the form of an open question argument, and that this "master argument" does a good job of capturing the real form of the underlying issues that pose obstacles for physicalist explanation of consciousness. In this section, I shall argue that putting the anti-physicalist arguments in this form actually helps to show that they are in a certain sense even stronger than their proponents have generally held them to be, and that seeing this helps to point the way to different explanatory issues and concerns than are current in contemporary discussions. In particular, whereas these arguments are standardly taken to establish that there cannot be a *physical* explanation of consciousness, I shall argue that restating them as open question arguments shows that they (like the arguments of Moore and Frege) can establish an even stronger result, namely that there cannot be (in a certain sense of "natural") any *naturalistic* explanation of consciousness at all. Here, moreover, "naturalistic" has the sense of a certain (logical) *type* of explanation, and so if the arguments are successful, they establish that we must look for an explanation of quite a different type, if we are to explain consciousness at all.

To begin to see this, we can note that the open question argument formulation is in fact exceedingly general. To make the open question argument, we considered whether a question about consciousness would still be open given any putative explanation of a certain type. In the first instance, we considered *physical* explanations, for instance explanations in terms of specific brain states. However, we can generate very similar (and equally plausible) arguments by considering putative explanations of other types, and arguing that given these other explanations there would still remain an open question. For instance, we can consider *functional* explanations. Just as there is plausibly an open question, given any putative explanation of consciousness in terms of physical brain states, as to *why* those physical brain states correlate (or are identical) with consciousness, there is plausibly an open question, given any putative explanation in terms of functional (or any dispositional) states, as to *why* those functional states correlate (or are identical) with consciousness. The zombie argument, in particular, captures this

very well, since zombies can be conceived as *functionally* identical with us but as still lacking phenomenal consciousness. Similarly, we can run the open question argument with respect to (any kind of) *causal* explanation of consciousness.[7] That is, given any explanation of the (non-phenomenal) processes or events that cause consciousness, it seems there will be an open question as to *why* these processes or events cause consciousness. This is also captured by the zombie scenario, since zombies are identical to us with respect to all non-phenomenal causal processes.

Noting the generality of this form of argument, as well as the similarity to Moore's and Frege's arguments against naturalism with respect to the good and truth, it is tempting (pending, of course, a definition of the "natural") to suggest that the open question argument bears against *any* "naturalist" attempt to explain consciousness whatsoever. If this were correct, it would *also* bear against a suggestion sometimes made even by some anti-physicalists about consciousness, the suggestion that (even if there is no physical explanation of consciousness), there might nevertheless be a "natural" explanation of another kind, for instance one that turns on the discovery of new laws or regularities in the natural world that are not simply "physical" in nature.

What is the reason, in the most general sense, for the irreducibility (or obstacle to reductive explanation) that yields the open question argument about consciousness for *all* of these possible types of explanation? Both in *The Conscious Mind* and elsewhere, Chalmers has argued (convincingly, in my opinion) that the deeper reason for the failure of both physical and functional explanations of consciousness has to do with the very nature of these kinds of explanations. In particular, both physical and functional explanations, like virtually all scientific explanations, explain what they explain in terms of *structure* and *function*. But no explanation purely in terms of structure and function suffices to explain consciousness. We can see this, in particular, by noting that, given any description of structures and functions (whether neurobiological, computational, chemical, or whatever) we can still raise the question: why should these structures and functions add up to consciousness? This is, of course, simply another version (perhaps the most general version) of the open question argument.

If this is right, it suggests that to confront the problem of explaining consciousness, we must consider fundamentally different *types* of explanation than those involved in the explanation of most (other) phenomena in the natural world. For (as Chalmers also argues) most phenomena in the natural world can be explained in terms of structures and functions. By contrast, consciousness appears to be fundamentally different, in that it resists explanation in terms of any number of structural or functional facts. How we should understand this resistance depends, of course, on what we understand to be involved in a "structural" or "functional" description. There are various accounts available, but a reasonable and general one is the following. A "structural" description of any domain of phenomena is one that characterizes those phenomena in such a way as to yield a Ramsey sentence correctly describing their existence and behavior. A Ramsey sentence is one that contains no constant terms for

[7] This apparently includes both interactionist dualist explanations and epiphenomenalist explanations on which consciousness is caused by (but does not cause) physical processes.

individuals, but rather implicitly defines all objects and processes in the domain in terms of their relational structure; for such a description, individual objects are simply the bearers of certain "roles," or what stand in certain relations to one another. (These roles are usually specified as *causal* roles, but there is no obvious reason why they need be causal rather than defined in terms of some other kind of relation). Such an explanation by way of Ramseyfication will apparently explain (almost) every physical event or process, but it will still not explain consciousness. For as Chalmers points out, given any Ramseyfied description of processes and structures, we still have the open question of why the occupants of the roles so defined should be *phenomenal* (rather than any other kind of) states. Moreover, given the zombie scenario, it is possible for there to be organisms for whom exactly the same Ramseyfied description holds as holds for us, but who are not conscious at all.

In *Philosophical History and the Problem of Consciousness*, I argued on historical grounds that the problem that Chalmers describes has the same logical (or meta-logical) form as one that Carnap and other members of the Vienna Circle already discussed. This is the problem of the relationship of conscious experience to the world of objective phenomena and processes of nature, what they conceived of as a logically and structurally unified whole amenable to objective description. In *The Logical Structure of the World*, Carnap argued that a necessary precondition for the objective explanation of any domain is the structuralization of that domain, i.e. the description of its objects and phenomena by means of structural definite descriptions. These are statements that contain no indexical, demonstrative, or tensed terms and contain no singular terms for individuals. Rather, they characterize the phenomena under consideration only as the bearers of certain relations, ultimately to be integrated into a total relational structure that captures the whole of the (objective) world. Still, Carnap thought that (at least for one version of the "structure" project) these relations might actually *be* relations among phenomenal experiences. The problem that became more and more urgent, and came to a head in the so-called "protocol sentence debate," was then the problem of just how to characterize this relationship between the "intrinsic" contents of experiences and the total structural whole of the world, which might (but apparently need not) be considered to consist of structures of these experiences standing in certain relations to one another.

If the problem of explaining consciousness is really a problem of the relationship of structure (in any of these senses) to content, then the most general form of open question argument can be formulated as follows:

1. Suppose consciousness could be explained in terms of some set of structurally defined properties and relations
2. Then, given that some state, c, satisfies these properties and relations, it would not be an open question whether c is a conscious state (or whether some organism in state c is conscious)
3. But given any set of structurally defined properties and relations, it is still an open question, given that c satisfies these properties and relations, whether c is a conscious state

4. Therefore consciousness cannot be explained in terms of any set of structurally defined properties and relations.

Putting the argument in this form has interesting consequences for the question of what types of explanation (if any!) of consciousness might succeed, even if physicalist explanations (as the more limited form of the open question argument holds) must fail. Both in *The Conscious Mind* and in other articles[8] Chalmers has suggested that, even if physicalist explanations of consciousness must fail, it might be possible nevertheless to explain consciousness in a broadly *naturalist* framework. One way to do this, he suggests, might be to find new laws or regularities that connect facts about phenomenal experience to physical facts, without reducing one sort of fact to the other. These new laws or regularities might fit in, Chalmers suggests, with general principles that connect the structure of phenomenal experience systematically to the structure of neurological processing or computation.

If the discovery of new laws or regularities of nature that are not basically physical in character might lead to an explanation of consciousness, then it might be possible to explain consciousness in a way that is still broadly "naturalistic" even if it is not physicalist or materialist. However, it seems to be a consequence of the general open question argument that it is in fact *not* possible to explain consciousness in this way. To see this, we can consider two ways in which the suggested new laws or regularities might work. First, they might be *fundamental* in character. That is, they might characterize the behavior of phenomena in the natural world at a basic level, as do the existing laws of physics. Alternatively, they might be non-fundamental: they might link the behavior of matter to that of phenomenal experience at a non-basic level. This is basically an *emergentist* position, whereby phenomenal experience comes about at a non-basic level, for instance when matter is combined or is functioning in a certain way. On either option, the new laws or regularities will predict that phenomenal experience of a certain type or configuration arises when certain other conditions, at least partially involving phenomena governed by existing physical laws, are satisfied.

However, the problem for both the fundamental and the non-fundamental option is that, on either option, it is apparently still possible to formulate an open question about consciousness *even given that the new laws or regularities obtain*. For instance, on the non-fundamental (emergentist) option, suppose that the new laws or regularities predict that a phenomenal experience of type A will arise when matter of a certain type or types is configured in a particular complex way. This linkage can then be expressed by means of structural definite descriptions, and it will be possible to write down a Ramsey sentence that expresses this linkage without involving any singular terms. There remains, then, an open question: *why* should phenomenal experience of a certain type arise when matter is configured in this way? Something similar holds on the fundamental option. Here, new fundamental laws predict that experiences (say, of types A_1 through A_n) arise under various (fundamental) conditions (say, $C_1 \ldots C_n$) where the description of these conditions also involves, in part, the description of physical matter and forces (otherwise there is no link between the new fundamental laws and the old ones). But all of these

[8] E.g. "Can We Construct a Science of Consciousness?" and "Consciousness and its Place in Nature"

regularities can themselves be written down structurally and expressed by means of Ramsey sentences. There remains an open question as to *why* any of them hold: why there is the link they express between the behavior of physical matter and that of phenomenal states.

To see this problem especially vividly, it is helpful to return to (a modified version of) the Mary case, which (as I have suggested) can be understood as formulating an open question argument. For this version (what we might call the "future Mary" case), we can imagine Mary growing up in her black-and-white room at some point *after* the discovery of all of the new regularities and laws that are purported to explain consciousness in a naturalistic (but not physicalist) way. That is, along with all of the "physical facts" that she is provided with on the standard Mary scenario, she is also informed of all the newly discovered facts, including newly discovered laws and regularities (either of a fundamental or non-fundamental kind), about the linkages between the behavior of physical matter and that of phenomenal states (described, e.g., as states of types $A_1 \ldots A_n$). The problem is that the situation of future Mary, when she leaves the black-and-white room, is *prima facie exactly like that of the original Mary*. Just like the original Mary, she is surprised to find out that *this is what it is like* to see blue; just like original Mary, she seems to learn something new, notwithstanding that she had already been informed about all the (fundamental or non-fundamental) laws governing phenomenal experience. If this is right, then it follows that even supplying Mary with all of the (physical as well as non-physical) *natural facts* does not suffice to give her knowledge of what it is like to see blue; therefore, we can conclude, there is no *naturalist* explanation (in terms of non-physical laws or regularities) of phenomenal experience.[9]

This helps us to see that the underlying problem here is not simply that the "physical facts" (with which Jackson's original thought experiment envisages Mary as being provided) are limited with respect to some broader category of structurally similar facts, but rather that there is a way in which phenomenal consciousness resists description and explanation in terms of *any* set or category of structurally expressible facts at all. This suggests that, although the thought experiment was originally put in terms of the extent of "physical facts" and is still routinely used to establish an anti-physicalist conclusion, the reference to "physical" facts is in fact superfluous and what is at issue is rather the limitations of a much broader *logical* category of facts of which the "physical facts" are a subset, what we may call the "structural" facts. These are facts that can be expressed objectively (i.e. in such a way that the meaning and truth value of the expression does not depend in any essential way on the circumstances of utterance or inscription) by means of sentences that make no irreducible reference to particulars and contain no indexical or demonstrative elements. If this is right, then it appears that the real underlying

[9] A proponent of non-physicalist naturalism might agree that all of the information that could be provided to Mary does not suffice to close the open question, while holding nevertheless that the right explanation for consciousness involves new laws or regularities that are broadly naturalistic, although Mary cannot understand them before she leaves the room. To hold this kind of position, however, is already to admit that the new laws or regularities that characterize consciousness in its relation to matter have a very different (semantic or logical) character than any of the other laws of nature, which *can* of course be stated objectively and understood by Mary on the basis of the textbooks and descriptions she has read before she leaves the room. It is not clear what is meant, then, by continuing to call these new regularities "naturalistic" at all.

problem with explaining consciousness is that the facts that we are attempting to explain do not have this form.

How far can we take this conclusion? Quite far indeed, it seems. For we can imagine Mary provided with *any* structural facts whatsoever – that is, any facts that can be written down in sentences that retain their meaning and truth value regardless of the context of their utterance or inscription – and it appears that she will still be in exactly the same situation when she leaves the room, *unless* she has already directly experienced the relevant phenomenal properties. In this way, the application of the general form of the open question argument appears to confirm that phenomenal properties, by contrast, are special in the sense that knowledge of them depends on their being (or instances of them being) *presented* to the knower. That this is not the case for (other) natural properties is, by contrast, precisely what the structuralizability of these other properties captures. And so the explanation of phenomenal consciousness, if it is to be successful, will have to consider the general *logical* issue of the relationship of structural explanation to "non-structural" facts such as those about phenomenal consciousness.

There is one more type of (broadly) "naturalist" explanation that should be considered as a possibility here (aside from the fundamental and emergentist types I considered above). This is the position of *intrinsic monism*, according to which the *non-relational* or 'intrinsic' properties of physical matter are ultimately phenomenal or proto-phenomenal, or perhaps (on a "dual-aspect" picture) are neutral between physicality and phenomenality but can be organized into phenomenal particulars. It might be thought that this kind of position offers the best response to the structuralist concerns about consciousness that we've discussed. For given that "facts about consciousness" apparently cannot be put in structural form, it is tempting to suppose that these facts may ultimately be *about* (what are actually) the *occupants* of the roles defined by a structural description of the natural facts (or the relevant Ramsey sentences). In particular, if these natural facts are exhausted by the physical facts, then we might suppose (as Russell seems to have suggested) that the (structuralist) description of the physicalist facts leaves out phenomenal (or proto-phenomenal) particulars *because* these are precisely the occupants of the roles defined by the structuralist description. Then we could try to supplement the physicalist/structuralist story, and complete the naturalist explanation of consciousness, by explaining that the occupants of these roles are precisely the phenomenal (or proto-phenomenal) particulars.

However, even though the neutral monist explanation takes into account the problem of structuralism and content, it is not clear that it can succeed as a naturalist explanation of consciousness either, as we can see by once again considering the open question argument. For just as there is an open question about consciousness even given any structuralist explanation, it seems there remains an open question even when we supplement the structuralist explanation with an identification of the structural role-bearers as phenomenal or proto-phenomenal properties in the way considered above. To see this, suppose a complete structural description of physical (or other natural) facts is given, for instance by means of a set of Ramsey sentences or a single, conjoined Ramsey "world-sentence", involving no singular terms but only variables $x_1 \ldots x_n$. This description is then supplemented by a series of

identifications of the variables in the structural description with phenomenal states or state-types. For instance, we supplement the structural description with the information that the existentially quantified variable x_i is *in fact* a state of experiencing phenomenal redness, etc.[10] The problem is that there still remains an open question: even given the supplemental information, why should *these* structural facts plus these identifying facts yield *these* phenomenal experiences? We can again put the problem in vivid form by considering the Mary case: even if Mary were provided with all the structural information *plus* the relevant information about the intrinsic identity of the role occupants in the structural description, she would still remain in the same epistemic situation upon exiting the room as was the case in the original Mary scenario. Given, in fact, *any information whatsoever* (that can be written down and whose expression does not essentially depend on context or involve direct reference to particulars), Mary will remain in the same epistemic situation – unless, of course, she somehow is able to *experience* the relevant phenomenal particulars herself.

III

On the basis of an extended application of the open question argument, I have argued that the problem of explaining consciousness is really a problem about the relationship between "structural facts" and facts or contents that cannot be expressed by structural descriptions, as almost all scientific facts can be. This means that the problem of explaining consciousness is really in many important respects a *logical* (or perhaps meta-logical) problem about the logical forms and structures of facts and (certain kinds of) particulars. Though this implication about the actual form of the problem has not generally come to the fore in the recent discussions, which have (somewhat misleadingly, if the current argument is right) presented the problem primarily as one of the limits of the "physical" facts, it is actually just the implication we should expect if we take a broader historical view. For, as I argued in *PHPC*, at the very beginning of the discussion of "physicalism" within the analytic tradition, the structuralism of Carnap and other early analytic philosophers already posed the problem of consciousness as the problem of the relationship of phenomenal contents to the objective and structuralizable descriptions of science.[11] This is, moreover, the problem that came to a head in the protocol sentence debate, when the question of the relationship of the structures of objective description to (what were supposed to be) their basis in experienced subjective contents became pressing. But although this question was raised and pursued in the course of the protocol sentence debate, it was not resolved therein or at any other subsequent stage of the analytic tradition, so it is not surprising that it remains a central and deep problem today.

If the argument I have made here is correct, then it has an implication which may seem to be "bad news" for certain kinds of explanatory projects: in particular, like the open question arguments of

[10] Or alternatively for the neutral-monist view, we can imagine being given two complete ramsifyed descriptions: one gives all the physical facts and one gives all the phenomenal facts, and then we are told that certain variables in one sentence correspond to (or range over individuals identical to) certain variables in the other. There still apparently remains an open question about *why* this should be so.

[11] This is partly because, in the Vienna Circle discussion, "physicalism" was simply identified with the view (held by *all* of the Vienna Circle prinicipals) that objective facts are structurally expressible within a structurally unified description of nature.

Moore and Frege, the current open question argument establishes the necessary failure of any attempt to explain the target phenomenon (here: consciousness) in (not only any physicalist but) any "naturalist" way at all. For it is very plausible that all facts and phenomena of the natural world are objectively expressible by means of structuralist descriptions, but as we have seen any such description or combination thereof appears to leave open questions about consciousness. As I have argued, this apparently means that not only physicalist, but even non-physicalist "naturalist" explanations (for instance "naturalistic dualist" ones) must fail, and accordingly that there is quite a different kind of problem to be confronted here.

On the other hand, though, there is a sense in which this is not after all bad news, for it points to the direction in which the question about consciousness might better be pursued. In particular, although on the current analysis the prospects for any kind of explanation of consciousness in terms of *facts about nature* appear to be dim, this does not mean that we must abandon all hope for any kind of scientific explanation at all. Quite to the contrary, we should consider that, although the current argument suggests that a *natural-scientific* explanation of consciousness will not be forthcoming, scientific knowledge is not limited to the natural sciences but includes the "formal sciences" of logic and mathematics as well. In particular, we should consider that where natural-scientific explanations are not forthcoming, it may nevertheless be possible to account for the relevant phenomena (i.e. to close the questions that appear to be left open by any natural-scientific description) in (broadly) *formal*-scientific terms.[12]

Seen in this light, the current anti-naturalist argument, to the effect that there remains an open question given no matter what natural-scientific (structural) facts or explanations, does not at all motivate eliminativism about consciousness, any more than Moore's argument motivates eliminativism about goodness or Frege's motivates eliminativism about truth. Rather, like these arguments, it suggests that the target phenomenon has a *logically* special character that distinguishes it from the phenomena of nature, and renders its expression in straightforward factual terms at least problematic. This suggests that the further analysis of consciousness, like reflection on goodness or truth on Frege's or Moore's analysis, should focus on this logically special character that distinguishes it from the phenomena treated by natural science, including those specific logical features or aspects that render it resistant to structural description and explanation.

In particular: as we have seen, if the general problem about consciousness is really a problem about the relationship between the structurally expressible facts and something that cannot be so expressed (whether or not this "something" can be put as consisting of *facts* at all) then the right way to resolve the problem is not to look for further natural-scientific or factual laws or regularities, but rather to clarify the issues of logical structure and form that lead to the persistence of the open question, given any number of structural, natural-scientific facts. It appears, in particular, that "facts about

12

consciousness" (or at any rate, what we are trying to explain in trying to explain consciousness) have a fundamentally different logical form than that of structurally expressible, natural-scientific facts.

What more can we say, though, about the nature of this apparent difference in logical form? In order to characterize it more closely, it will be important to clarify exactly *why* consciousness is different than virtually all other scientifically describable phenomena (except, perhaps, phenomena like goodness and truth, which are of course the objects of the other open question arguments we've considered) in that there remain open questions about it even when all structurally expressible facts are given. Actually, it is not difficult to see at least the outlines of an answer. As we have seen repeatedly, both in connection with the Mary case and with Kripke's own analysis of reference to phenomenal states, phenomenal states appear to have the peculiarity that it is *necessary* (if not sufficient), in order for some facts about them to be known, that *they themselves* be *presented to* or *experienced by* the knower. This is, by contrast, *not* the case for any familiar type of natural fact or any phenomenon describable in purely structural terms. This is plausibly the reason why questions about consciousness remain open, as in the Mary case, given any amount of structural information (physical or non-physical), as long as she has not in fact experienced the relevant phenomenal states herself. And although it may appear to be a primarily epistemological difference (since it is put in terms of what is necessary for knowledge), what it really gestures to is not an epistemological but rather a logical or "metalogical" point: that phenomenal states and their characters are, in a certain essential sense, *self*-presentational, and in such a way that they essentially *cannot* be presented (in a way sufficient to confer knowledge) by any objective, structural description.[13]

This helps, as well, to suggest elements of existing analyses that might be drawn upon in developing a better analysis of the underlying logical issues here. For instance, as we have seen, it is plausible that the open question about phenomenal content remains for Mary given any information whatsoever that can be written down in a structural and context-independent form. Of course, there are other important types of information (or aspects of our information) about the world that *cannot* be written down in this kind of way, and yet are still essential to our orientation and navigation. An obvious example is "indexical information", such as what is expressed by means of (tokens of) sentences such as "It is now 5 o'clock" or "I am currently in Canberra." This kind of information is at least analogous to phenomenal information in that it cannot be straightforwardly translated into structural descriptions, and so is in this sense "above and beyond" the information available to Mary before she steps out of the black-and-white room. Based on this, John Perry has suggested that the problem of phenomenal consciousness might be solved (in a way friendly to physicalism) by treating Mary's phenomenal

[13] This is not to say simply that phenomenal states are "ineffable" or inexpressible in "public language." For, of course, *given* that a subject has knowledge of what it is like, e.g., to see blue, it is perfectly possible for that subject to understand another's description of her current phenomenal state as one of experiencing phenomenal blueness. Still, it appears to be an implication of the current analysis that there is a "deep" interconnection between issues about the structure of consciousness and issues of objective, public and intersubjective expressibility, which can perhaps be put by noting that there is a link between structural expressibility and expressibility in a *general* (or "public") sense, i.e. in a way that does not depend on the context of expression or on the previous experiences of the listeners or readers.

discovery upon stepping out of the room as having the same form as her discovery of an "essential indexical" (for instance: *That* guy is Fred Dretske.) However, although there is certainly an *analogy* here between phenomenal information and "indexical information," (in that both escape expression in structuralist form) it is not clear that this is any more than an analogy, or (as has been pointed out against Perry) that it implies that gaining phenomenal knowledge involves gaining indexical knowledge in any essential way. (It is in fact not even clear that there is any such distinctive sort of knowledge as the latter, or that if there is, it can indeed be wholly explained in physical terms, as Perry assumes). Further work on this connection should explore more closely the deeper phenomena of logical structure, reflexivity (token- or otherwise), and presentation that appear to underlie this analogy in order to get clearer about the underlying logical phenomena here.

Again, as we have seen it is plausible that the logical peculiarity of phenomenal states that we have noted is a result of (or at any rate is connected to) the fact that these states are in a distinctive way self-presentational, so that it is actually *necessary* in order to have knowledge of their properties that they actually be presented to us. This suggests connections to what has been called the "phenomenal concept strategy," which is an increasingly popular way of attempting to explain at least the appearance of an explanatory gap between physical concepts and phenomenal ones. On at least some versions of this strategy, the appearance of a gap arises because (although it is the same states that are presented both physically and phenomenally) it is necessary for the fixation of our phenomenal concepts that the states falling under those concepts themselves be presented. As I have argued elsewhere, it is in fact not at all clear that these analyses are actually consistent with physicalism at all, especially since many of them rely on an appeal to a distinctive kind of "cognitive" presentation of content that closely resembles what was classically discussed as "acquaintance" and cannot obviously be analyzed in purely physicalist terms.[14] Nevertheless, it is probably worthwhile to pursue various versions of the "phenomenal concept strategy" (both physicalist and non-) in order to gain more insight into the distinctive logical features of phenomenal states and their concepts that mark both off from other sorts of states and concepts.

Finally, the idea of a logically distinctive "self-presentational" status of consciousness, which may partially account for the persistence of open questions that we've noted here, obviously suggests connections to "self-representational" or single-order reflexive theories of consciousness (such as that suggested by Kriegel (2009)). Although, on the present analyses, these theories should perhaps be better treated as "self-presentational" rather than "self-*re*presentational" theories of conscious states (so as not to presuppose that a theory of intentional *representation* is a necessary preliminary for any account of the special logical status of consciousness in these respects), it is certainly worth pursuing further the formally and logically specific phenomena of reflexivity that these theories suggest may essentially characterize consciousness. In particular, it is to be hoped that a more detailed pursuit of the formally peculiar presentational features of consciousness, in connection with a continuation of the

[14] "Phenomenal Concepts and the Problem of Acquaintance"

classic meta-logical inquiry into the consequences of reflexivity, may help to clarify the logical and meta-logical status of the phenomena at issue here.