

Phenomenal Concepts and the Knowledge Argument

[David J. Chalmers](#)

Department of Philosophy
University of Arizona
Tucson, AZ 85721.

chalmers@arizona.edu

*[[This paper is largely based on material in other papers. The first three sections and the appendix are drawn with minor modifications from Chalmers 2002c (which explores issues about phenomenal concepts and beliefs in much more depth, mostly independently of questions about materialism). The main ideas of the last three sections are drawn from Chalmers 1996, 1999, and 2002a, although with considerable revision and elaboration.]]

1 Introduction

The classic statement of the knowledge argument against materialism has been given by Frank Jackson (1982). Mary knows everything that can be stated in physical terms about the physical processes that are in any way relevant to color vision. But Mary has never experienced colors other than black, white, and shades of grey. It seems that Mary has complete physical knowledge, but she does not have complete phenomenal knowledge: in particular, she does not know what it is like to see red. Jackson argues that Mary knows all the physical facts, but does not know all the facts. When she sees red for the first time, she learns a new fact concerning what it is like to see red. So there are facts over and above the physical facts, and materialism is false. In particular, phenomenal facts — facts about the character of conscious experience — are nonphysical facts, and phenomenal properties are nonphysical properties.

The knowledge argument turns crucially on Mary's new knowledge, and on her acquisition of new beliefs. To understand the nature of this knowledge, we need to understand the concepts — *phenomenal concepts* — that are involved in Mary's new beliefs. In this paper, I will give an analysis of these concepts (in sections 2 and 3), and will then bring it to bear on the knowledge argument itself (in sections 4 and later). I will argue that the knowledge argument is basically sound (section 5), and that a correct understanding of phenomenal concepts helps to see why many responses to the knowledge argument fail (section 6).

In what follows, I will assume *phenomenal realism*: roughly, the view that Mary acquires new factual knowledge (not a priori deducible from physical knowledge) when she sees red for the first time. This excludes views on which Mary merely gains a new ability, or on which she gains no knowledge at all. It is compatible with views on which Mary gains knowledge of an old fact in a new way. The important aspect of this view is that it allows an *epistemic* gap between physical truths and phenomenal truths, in the sense that phenomenal truths are not entailed a priori by physical truths. The view is silent on whether or not there is an ontological gap. As such, the view excludes "type-A" materialist views such as those of Dennett and Lewis, which deny an epistemic gap, but it includes "type-B" materialist views such as those of Loar and Tye, which allow an epistemic gap but denying an ontological gap, as well as including many non-materialist views.

2 Phenomenal concepts

Phenomenal properties are properties characterizing what it is like to be a subject, or what it is like to be in a mental state. Phenomenal beliefs are beliefs that attribute phenomenal properties. I will be especially concerned with first-person phenomenal beliefs, such as *I am now having a red experience*. Phenomenal

beliefs attribute phenomenal properties using phenomenal concepts. To understand phenomenal beliefs, we need to understand phenomenal concepts.

(In this paper, I understand beliefs and concepts as psychological entities rather than as semantic entities. Beliefs and concepts have contents, but are not themselves contents.)

Mary looks at a red apple, and visually experiences its color. This experience instantiates a phenomenal property R, which we might call phenomenal redness. It is natural to say that Mary is having a red experience, even though of course experiences are not red in the same sense in which apples are red. Phenomenal redness (a property of experiences, or of subjects of experience) is a different property from external redness (a property of external objects), but both are respectable properties in their own right.

Mary attends to her visual experience, and thinks *I am having an experience of such-and-such quality*, referring to the quality of phenomenal redness. There are various concepts of the quality in question that might yield a true belief.

We can first consider the concept expressed by 'red' in the public-language expression 'red experience', or the concept expressed by the public-language expression 'phenomenal redness'. The reference of these expressions is fixed via a relation to red things in the external world, and ultimately via a relation to certain paradigmatic red objects that are ostended in learning the public-language term 'red'. A language learner learns to call the experiences typically brought about by these objects 'red' (in the phenomenal sense), and to call the objects that typically bring about those experiences 'red' (in the external sense). So the phenomenal concept involved here is *relational*, in that it has its reference fixed by a relation to external objects. The property that is referred to need not be relational, however. The phenomenal concept plausibly designates an intrinsic property rigidly, so that there are counterfactual worlds in which red experiences are never caused by red things.

One can distinguish at least two relational phenomenal concepts, depending on whether reference is fixed by relations across a whole community of subjects, or by relations restricted to the subject in question. The first is what we can call the *community relational concept*, or red_C . This can be glossed roughly as *the phenomenal quality typically caused in normal subjects within my community by paradigmatic red things*. The second is what we can call the *individual relational concept*, or red_I . This can be glossed roughly as *the phenomenal quality typically caused in me by paradigmatic red things*. The two concepts red_C and red_I will corefer for normal subjects, but for abnormal subjects they may yield different results. For example, a red/green-inverted subject's concept red_C will refer to (what others call) phenomenal redness, but his or her concept red_I will refer to (what others call) phenomenal greenness.

Phenomenal properties can also be picked out indexically. When seeing the tomato, Mary can refer indexically to a visual quality associated with it, by saying 'this quality' or 'this sort of experience'. These expressions express a demonstrative concept that we might call *E*. *E* functions in an indexical manner, roughly by picking out whatever quality the subject is currently ostending. Like other demonstratives, it has a "character", which fixes reference in a context roughly by picking out whatever quality is ostended in that context; and it has a distinct "content", corresponding to the quality that is actually ostended — in this case, phenomenal redness. The demonstrative concept *E* rigidly designates its referent, so that it picks out the quality in question even in counterfactual worlds in which no-one is ostending the quality.

The three concepts red_C , red_I , and *E* may all refer to the same quality, phenomenal redness. All of them fix reference to phenomenal redness relationally, characterizing it in terms of its relations to external objects or acts of ostension, and all of them designate this quality rigidly.

There is another crucial phenomenal concept in the vicinity, one that does not pick out phenomenal redness in terms of its relation to external objects or to acts of ostension, but rather picks it out in terms of its intrinsic phenomenal nature. This is what we might call a *pure phenomenal concept*.

To see the need for the pure phenomenal concept, consider the knowledge that Mary is in a position to gain when she learns for the first time what it is like to see red. She may learn that seeing red has such-and-such quality. Mary may learn (or reasonably come to believe) that red things will typically cause experiences of such-and-such quality in her, and in other members of her community. She may learn (or gains the cognitively significant belief) that the experience she is now having has such-and-such quality, and that the quality she is now ostending is such-and-such. Call Mary's "such-and-such" concept here R .

Mary's concept R picks out phenomenal redness, but it is quite distinct from the concepts red_C , red_I , and E . We can see this by using cognitive significance as a test for difference between concepts. Mary is in a position to gain the belief $red_C=R$ — that the quality typically caused in her community by red things is such-and-such — and this belief is cognitively significant knowledge. She may gain the cognitively significant belief $red_I=R$ in a similar way. And she may gain the belief $E=R$ — roughly, that the belief that *this* quality is such-and-such.

Mary's belief $E=R$ is as cognitively significant as any other belief in which the object of a demonstrative is independently characterized: e.g. my belief *I am David Chalmers*, or my belief *that object is tall*, or my belief *that shape is roundness*. For Mary, $E=R$ is not a priori. No a priori reasoning can rule out the hypothesis that some other quality is the object of her current ostension, just as no a priori reasoning can rule out the hypothesis that I am David Hume, or that the object I am pointing to is short. Indeed, nothing known a priori entails that R is ever instantiated in the actual world.

So the concept R is quite distinct from red_C , red_I , and E . Unlike the other concepts, the pure phenomenal concept picks out the phenomenal quality *as* the phenomenal quality that it is.

(The distinction between pure and relational phenomenal concepts roughly tracks Nida-Rümelin's (1995) distinction between "phenomenal" and "nonphenomenal" readings of belief attributions concerning phenomenal states. "Phenomenal" belief attributions can only be satisfied by beliefs involving pure phenomenal concepts, while "nonphenomenal" belief attributions can be satisfied by beliefs involving either pure or relational phenomenal concepts.)

It may be that there is a broad sense in which R can be regarded as a "demonstrative" concept. I will not regard it this way: I take it that demonstrative concepts work roughly as analyzed by Kaplan (1977), so that they have an reference-fixing "character" that leaves their referent open. This is how E behaves: its content might be expressed roughly as "this quality, whatever it happens to be". R , on the other hand, is a substantive concept that is tied *a priori* to a specific sort of quality, so it does not behave the way that Kaplan suggests that a demonstrative should. Still, there is an intimate relationship between pure and demonstrative phenomenal concepts that I will discuss later in the paper; and if someone wants to count pure phenomenal concepts as "demonstrative" in a broad sense, there is no great harm in doing so, as long as the relevant distinctions are kept clear. What matters for my purposes is not the terminological point, but the more basic point that the distinct concepts E and R exist.

3 The content of phenomenal concepts

The relations among these concepts can be analyzed straightforwardly using the two-dimensional framework for representing the content of concepts. A quick introduction to this framework is given in an appendix; more details can be found in Chalmers (2002b).

According to the two-dimensional framework, when an identity $A=B$ is a posteriori, the concepts A and B have different epistemic (or primary) intensions across epistemically possible scenarios.[*] If A and B are rigid concepts and the identity is true, A and B have the same subjunctive (or secondary) intensions. So we should expect that the concepts red_C , red_I , E , and R have different epistemic intensions, but the same subjunctive intension. And this is what we find. The subjunctive intension of each picks out phenomenal redness in all worlds. The epistemic intension of red_C picks out, in a given centered world, roughly the quality typically caused by certain paradigmatic objects in the community of the subject at the center of the

world. The epistemic intension of red_I picks out roughly the quality typically caused by those objects in the subject at the center.

(Note that for the purposes of analyzing the content of phenomenal concepts, epistemic intensions should be taken as functions across epistemically possible scenarios or conceivable worlds, without assuming that these must coincide with metaphysically possible worlds.)

As for the demonstrative concept E : to a first approximation, one might hold that its epistemic intension picks out the quality that is ostended by the subject at the center. This characterization is good enough for most of our purposes, but it is not quite correct. It is possible to ostend two experiences simultaneously and invoke two distinct demonstrative concepts, as when one thinks *that quality differs from that quality*, ostending two phenomenal color qualities associated with symmetrical spatial locations in a symmetrical visual experience (see Austin 1990). Here no descriptive characterization such as the one above will capture the difference between the two concepts. It is better to see E as a sort of indexical, like I or now (hence the name E , for a primitive experiential indexical). To characterize the epistemic possibilities relevant to demonstrative phenomenal concepts, we need centered worlds whose centers contain not only a "marked" subject and time, but also one or more marked experiences; in the general case, a sequence of such experiences. Then a concept such as E will map a centered world to the quality of the "marked" experience (if any) in that world. Where two demonstrative concepts $E1$ and $E2$ are involved, as above, the relevant epistemic possibilities will contain at least two marked experiences, and we can see $E1$ as picking out the quality of the first marked experience in a centered world, and $E2$ as picking out the quality of the second. Then the belief above will endorse all worlds at which the quality of the first marked experience differs from the quality of the second. This subtlety will not be central for the purposes of this paper.

The epistemic intension of R is quite distinct from all of these. It picks out phenomenal redness in all worlds. When Mary believes *roses cause R experiences*, or *I am currently having an R experience*, she thereby excludes all epistemic possibilities in which roses cause some other quality (such as G , phenomenal greenness), or in which she is experiencing some other quality: only epistemic possibilities involving phenomenal redness remain.

The cognitive significance of identities such as $red_C=R$, $red_I=R$, and $E=R$ is reflected in the differences between the concept's epistemic intensions. The first two identities endorse all epistemic possibilities in which paradigmatic objects stand in the right relation to experiences of R ; these are only a subset of the epistemic possibilities available a priori. The third identity endorses all epistemic possibilities in which the marked experience at the center (or the ostended experience, on the rough characterization) is R . Again, there are many epistemic possibilities (a priori) that are not like this: centered worlds in which the marked experience is G , for example. Once again, this epistemic contingency reflects the cognitive significance of the identity.

(The rest of this section is inessential for the discussion of the knowledge argument in the rest of this paper, but is independently relevant to the analysis of phenomenal concepts and phenomenal knowledge in Mary's situation. As with the previous material, the material that follows is developed at greater length in Chalmers 2002.)

One could even consider the conceivable case of *Inverted Mary*, who is physically, functionally, and environmentally just like Mary, except that her phenomenal color vision is red/green inverted. Like Mary, Inverted Mary learns something new when she sees red things for the first time. But Inverted Mary learns something different from what Mary learns. Where Mary learns that tomatoes cause experiences of (what we call) phenomenal redness, Inverted Mary learns that they cause experiences of (what we call) phenomenal greenness. In the terms given earlier, Mary acquires beliefs $red_C=R$, $red_I=R$, and $E=R$, while Inverted Mary acquires beliefs $red_C=G$, $red_I=G$, and $E=G$ (where G is the obvious analog of R). So Mary and Inverted Mary acquire beliefs with quite different contents.

Even after they see red things for the first time, Mary and Inverted Mary are physical and functional twins. Nevertheless, they have beliefs with different contents. It follows that belief content does not supervene conceptually on physical/functional properties. That is: once we grant that phenomenal properties are conceptually irreducible to physical/functional properties, we must grant that the same goes for at least some intentional properties.

This case might seem analogous to Putnam's familiar case of Twin Earth (Putnam 1975), but something different is going on. On Twin Earth, where the water in our environment is replaced by the superficially identical but chemically distinct XYZ, when Twin Oscar says "water is wet", he expresses a belief that differs in content from the corresponding belief expressed by Oscar on Earth. Oscar's belief is about XYZ, where Twin Oscar's belief is about H₂O. But one can argue that while these beliefs differ in their subjunctive content, their epistemic content is the same: roughly, the epistemic content of both beliefs endorses all epistemic possibilities in which the watery stuff in the environment of the subject at the center is wet. That is, Oscar's and Twin Oscar's "water" concepts have different extensions, and different subjunctive intensions, but they have the same epistemic intension.

By contrast, Mary's concept *R* and Inverted Mary's concept *G* differ not only in their extension and in their subjunctive intension, but also in their epistemic intension. When Mary thinks *I am having an R experiences*, the epistemic content of her belief endorses only those epistemic possibilities in which the subject at the center has an R experience. When Inverted Mary thinks *I am having a G experience*, the epistemic content of her belief endorses only those epistemic possibilities in which the subject at the center has a G experience. So unlike the Twin Earth cases, this appears to be a case in which the epistemic content of a subject's belief does not supervene conceptually on the subject's physical and functional properties.

Something unusual is going on here. In standard externalism, and in standard cases of so-called "direct reference", a referent plays a role in constituting the subjunctive content (subjunctive intension) of concepts and beliefs, while leaving the epistemic content (epistemic intension) unaffected. In the pure phenomenal case, by contrast, the quality of the experiences plays a role in constituting the *epistemic* content of the concept and of the corresponding belief. One might say very loosely that in this case, the referent of the concept is somehow present inside the concept's sense, in a way much stronger than in the usual cases of "direct reference".

One might say that concepts such as *water* are *subjunctively rigid*, in that their subjunctive intensions pick out the same extension in all possible worlds. Pure phenomenal concepts, by contrast, are not only subjunctively rigid but *epistemically rigid*, in that their epistemic intensions pick out the same extension in all scenarios. Martine Nida-Rümelin suggests in a forthcoming paper that pure phenomenal concepts can be called *super-rigid*: when represented as a two-dimensional matrix, they pick out the same extension in all locations of the matrix. Super-rigidity is clearly a much stronger phenomenon than standard rigidity.

4 Phenomenal Concepts and the Knowledge Argument

This analysis of phenomenal concepts bears on the knowledge argument in a number of ways. For a start, it gives us a better characterization of Mary's new knowledge. The crucial new beliefs that Mary gains do not involve just any phenomenal concept that refers to phenomenal redness, but rather involve a pure phenomenal concept such as *R*. These beliefs have the form *this experience is R, I am experiencing R, red things typically cause R experiences, other people experience R when they look at tomatoes, R is instantiated*, and so on. The content of *R* is tied to the phenomenal property R in a very direct way: both the epistemic and subjunctive intensions of the concept pick out instances of *R* in all possible worlds. And this content appears to be determined directly by the instantiation of R itself in Mary: if Inverted Mary instantiates G instead, she will have a quite different pure phenomenal concept *G*, and quite different resulting beliefs.

Nevertheless, what I have said so far is already enough to see that certain materialist *responses* to the knowledge argument fail. It is common for materialists to respond to the knowledge argument by invoking

specific claims about the nature of phenomenal concepts, or analogies with other concepts. For example, some (e.g. Ismael 1999, Perry 2001) hold that phenomenal concepts are *indexical* concepts, so that the epistemic gap here can be assimilated to the epistemic gap between objective knowledge and indexical knowledge more generally. Others (e.g. Hawthorne 2002, Loar 1997) hold that phenomenal concepts are *demonstrative* or *recognitional* concepts, so that the epistemic gap can be assimilated to that that holds between theoretical knowledge and demonstrative or recognitional knowledge more generally.

From the discussion above, it is clear that Mary's crucial phenomenal concept is not an indexical or a demonstrative concept. Mary does have a concept of this form: it is her demonstrative phenomenal concept *E*. That concept behaves roughly as the accounts above suggest that phenomenal concepts should. But Mary's important new knowledge involves not *E* but *R*, and *R* is not an indexical or demonstrative concept at all. Rather, it is a pure phenomenal concept, a concept characterizing the phenomenal property in question directly in terms of its phenomenal character.

One can also make a direct case against any analysis of phenomenal knowledge as indexical or demonstrative knowledge, as follows. In the indexical case, any epistemic gaps disappear from an objective perspective. Say that I am physically omniscient, but do not know whether I am in the USA or Australia (let's imagine that there are appropriate qualitative twins in both). Then I have a certain indexical ignorance, and discovering that I am in the USA will constitute new knowledge. But if someone else is watching the world from the third-person point of view and is also physically omniscient, they will have no corresponding ignorance: they will know that A is in Australia and that B is in the USA, and that is all there is for them to know. That is, there is no thought about my location about whose truth-value they are ignorant: they know everything there is to know about my situation. So my ignorance is *essentially* indexical, and evaporates from the objective viewpoint. The same goes for indexical ignorance concerning what time it is, for demonstrative ignorance concerning what *this* is, and so on. In all these cases, the ignorance disappears from the objective viewpoint: an objectively omniscient observer can know everything there is for them to know about my situation, and there will be no doubts for them to settle.

Now consider Mary's ignorance. From her black-and-white room, she is ignorant of all sorts of facts: what it will be like for her to see red for the first time, what it is like for others to see red, and so on. Only the first of these looks even apparently indexical, so let us focus on that. In this case, a physically omniscient observer may have precisely analogous ignorance: even given his complete physical knowledge, he can entertain the question of what it is like for Mary to see red for the first time, and may have no idea what the answer is. So this ignorance does not evaporate from the objective viewpoint. The same goes even more strongly for knowledge of what it is like for others to see red. For any observer, regardless of their viewpoint, there will be an epistemic gap between complete physical knowledge and this sort of phenomenal knowledge. This suggests very strongly that phenomenal knowledge is not a variety of indexical or demonstrative knowledge at all. Rather, it is a sort of non-indexical knowledge of the world, not essentially tied to any viewpoint.

If this is right, then any analysis of phenomenal concepts as indexical or demonstrative concepts fails, and any attempt to explain Mary's epistemic gap in terms of the epistemic gap for indexical or demonstrative concepts fails.

There is more to say here. But to explore these issues, it is useful to first set out the two-dimensional analysis of the knowledge argument.

5 The Two-Dimensional Analysis of the Knowledge Argument

Nothing I have said so far entails that the knowledge argument is sound. So far, what I have said can be embraced in principle by a type-B materialist who holds that phenomenal properties are identical to physical properties, but that phenomenal concepts are distinct from physical concepts. The type-B materialist can take on board everything so far as an epistemic point about the distinctive behavior of phenomenal concepts, and hold that no non-materialist ontological consequences follow.

In my view, the knowledge argument is strongest when it is conjoined with the two-dimensional semantic framework, which allows us to think about the connection between epistemic and ontological matters, and between concepts and properties, in a more precise way. Once this is done, we can also bring in the analysis of phenomenal concepts given above, to help see why certain materialist *responses* to the knowledge argument fail.

The basic intuition arising from the knowledge argument is that Mary gains new factual knowledge when she sees red for the first time, knowledge that no amount of physical information and a priori reasoning would have allowed her to possess beforehand. Let P be the complete microphysical truth about the world, and let Q be a truth stating that phenomenal redness is instantiated, deploying a pure phenomenal concept of phenomenal redness. Then the initial moral of the knowledge argument is that Q cannot be deduced from P by a priori reasoning. That is, the material conditional 'P \square Q' is not knowable a priori.

This is an epistemic claim, and as such does not immediately suffice to refute the metaphysical thesis of physicalism. To do this, one needs a bridge from the epistemic claim to a metaphysical claim. As often in philosophy, this can be done by proceeding first from an epistemic claim to a modal claim (about necessity and possibility), and from there to a metaphysical claim. Here, the most straightforward version of such an argument would be the following:

- (1) 'P \square Q' is a posteriori.
- (2) If 'P \square Q' is a posteriori, 'P \square Q' is contingent.
- (3) If 'P \square Q' is contingent, physicalism is false.

—

- (4) Physicalism is false.

Here, premise (1) is a version of the epistemic claim above., Premise (2) is an instance of a traditionally popular thesis relating the epistemic and the modal, holding that when a sentence S is a posteriori, S is contingent. Premise (3) states a modal constraint on the metaphysical thesis of physicalism.

Premise (3) is intended to capture the general idea that if things could have been physically just as they are in our world but mentally different, then physicalism is false in our world. As such, the form of (3) needs minor modification: the truth of physicalism in our world seems compatible with the possibility of physically identical worlds with *additional* nonphysical minds. But this problem can be solved straightforwardly by conjoining to P a "that's-all" claim T, saying that our world is a *minimal* world satisfying P (roughly, a world containing no more than it needs to in order to satisfy P). If we replace P by the conjunction PT, premise (3) states a plausible constraint on physicalism (given that Q is true in our world), and the plausibility of the other premises is unchanged.

A more serious problem is premise (2) is widely believed to be false. Since Kripke (1980), many have accepted that some statements are both a posteriori and necessary: 'water is H₂O', for example, and 'Hesperus is Phosphorus'. If this is right, then it appears that there is no good reasons to accept premise (2). So the key premise connecting epistemic and modal claims is undercut, throwing doubt on arguments (such as the knowledge argument) that attempt to draw metaphysical conclusions from epistemic premises.

Here, the two-dimensional semantic framework becomes relevant. This framework allows us to draw a slightly different connection between the epistemic and modal domains: one that is compatible with the existence of necessary a posteriori statements, but that still allows us to draw metaphysical conclusions from epistemic premises.

Let us say that a sentence S is 1-necessary when its epistemic intension is true at all centered metaphysically possible worlds, and that it is 1-contingent when its epistemic intension is false at some centered metaphysically possible world. Let us also say that a sentence S is 2-necessary when its subjunctive intension is true at all worlds, and that it is 2-contingent when its subjunctive intension is false at some world. Then the following claim is crucial:

2D Thesis: If S is a posteriori, S is 1-contingent.

This thesis is plausibly true of all the a posteriori necessary statements that Kripke considers. For example, the epistemic intension of 'water is H₂O' is false at a Twin Earth centered world. The epistemic intension of 'Hesperus is Phosphorus' is false at a centered world where the evening star near the center is distinct from the morning star near the center. And so on. All these worlds are metaphysically possible. The claims above are quite compatible with Kripke's claim that these sentences are necessary. In effect, Kripke's claim is that the *subjunctive* intension of these sentences are true in all worlds, or that they are 2-necessary. This is quite compatible with their *epistemic* intensions being false in some worlds.

The 2D thesis above allows us to make inferences from epistemic claims to claims about metaphysical possibility, and from there to metaphysical conclusions. As such the thesis is substantive rather than trivial, and we will look later at attempts to deny it. For now, it is enough to note that the principle appears to fit all of Kripke's cases.

(A related thesis holds that when S is a posteriori, its epistemic intension is false at some epistemically possible scenario. This purely epistemic thesis, by contrast to the last, is more or less trivial on the two-dimensional framework, but does not license inferences from epistemic claims to metaphysical conclusions. In what follows, it will always be metaphysically possible worlds rather than epistemically possible scenarios that are relevant.)

With the 2D thesis in hand, we can reformulate the version of the knowledge argument above:

- (1) 'PT \square Q' is a posteriori.
- (2) If 'PT \square Q' is a posteriori, 'PT \square Q' is 1-contingent.
- (3) If 'PT \square Q' is 1-contingent, 'PT \square Q' is 2-contingent.
- (4) If 'PT \square Q' is 2-contingent, physicalism is false.

—

- (5) Physicalism is false.

Here, (1) is the epistemic thesis arising from the Mary situation, (2) is an instance of the 2D thesis, and (4) is a version of the modified modal constraint on physicalism, discussed above. Note that it is 2-necessity that is plausibly relevant to physicalism: physicalism requires that it *could not have been* that things were physically just as they are in our world (with nothing more), but mentally distinct. This is a subjunctive rather than an epistemic claim, so that 2-contingency rather than 1-contingency is directly relevant.

The remaining thesis is premise (3), bridging 1-contingency and 2-contingency. It is not true in general that 1-contingent statements are 2-contingent: counterexamples include 'water is H₂O', and 'Hesperus is Phosphorus'. The reason is that expressions such as 'water' and 'Hesperus' have quite different epistemic and subjunctive intensions. However, the principle is true for statements including only *semantically neutral* expressions, whose epistemic intensions are the same as their subjunctive intensions (that is, the

epistemic intension's value at any centered world is the same as the subjunctive intension's value at the corresponding uncentered world).

If P, T, and Q were semantically neutral, premise (3) would be true. It is plausible that Q is in fact semantically neutral, as we saw previously that pure phenomenal concepts have the same epistemic and subjunctive intensions, so that Q is semantically neutral. However, it is arguable that P is not semantically neutral. It is plausible that terms for microphysical properties, such as 'charge', refer rigidly to intrinsic properties, but pick out those properties by virtue of the fact that they play a certain causal role in our world (e.g. a certain role in electromagnetic processes). If so, then at any given world, the epistemic intension of 'charge' picks out whatever property plays the relevant causal role in the world, while the subjunctive intension picks out the intrinsic property (charge) that plays the causal role in *our* world. And it is arguable that these intensions differ, since there are arguably worlds where the relevant causal role is played by a property distinct from the property playing the role in our world. If so, premise (3) is false.

However, this opens up only a small loophole in the argument. For premise (3) to be false, there must be a world in which the epistemic intension of 'PT \square Q' is false, but in which its subjunctive intension is true. This must be a world in which the epistemic intension (and subjunctive intension) of Q is false, in which the epistemic intension of PT is true, and in which the subjunctive intension of PT is false. It is not hard to see that this must be a world in which things are structurally just as they are in our world, but with a different array of intrinsic properties playing the causal role that intrinsic microphysical properties play in our world. And this difference in intrinsic properties is responsible for the difference in truth-value of Q between that world and ours. That is, the causal structure of our microphysical world does not necessitate Q, but the intrinsic properties underlying that structure do (perhaps in conjunction with the structure). This is a version of the Russell-inspired view that I have called *panprotopsychism*.

We can therefore say: if 'PT \square Q' is 1-contingent but not 2-contingent, then panprotopsychism is true. Or equivalently: if 'PT \square Q' is 1-contingent, then 'PT \square Q' is 2-contingent or panprotopsychism is true.

This allows us to formulate the final version of the argument:

- (1) 'PT \square Q' is a posteriori.
- (2) If 'PT \square Q' is a posteriori, 'PT \square Q' is 1-contingent.
- (3) If 'PT \square Q' is 1-contingent, 'PT \square Q' is 2-contingent or panprotopsychism is true.
- (4) If 'PT \square Q' is 2-contingent, physicalism is false.

—

- (5) Physicalism is false or panprotopsychism is true.

Here, (1) is the epistemic claim arising from the Mary situation, (2) is an instance of the 2D thesis, (3) is the result of a straightforward piece of reasoning, while (4) is the modal constraint on physicalism. It is not clear whether (5) is as strong as a denial of physicalism, since it is not clear whether or not panprotopsychism is a form of physicalism. But if it is a form of physicalism, it is clearly a strange and unusual form, so the conclusion of the argument remains strong either way.

So here we have a very promising version of the knowledge argument: a valid argument for a strong ontological conclusion about consciousness, based on the epistemic intuition about the Mary case along with three other independently plausible premises.

6 Responses to the knowledge argument

We can use the argument above to analyze various responses that have been made to the knowledge argument.

(i) *The ability reply*: According to this type-A response (Lewis 1991, Nemirow 1991), Mary does not gain new factual knowledge, but merely gains an ability. Proponents of this response will deny that there are phenomenal truths that Mary cannot know in her room, and so will deny either premise (1) or the claim that Q is a truth. The same goes for other positions (e.g. Dennett 1991) according to which Mary gains no factual knowledge. The analysis above has no special force against this position, as the discussion here takes Mary's new factual knowledge for granted. Nevertheless, I think there are good reasons to reject the analysis (see e.g. Loar 1997; Nida-Rümelin 1995).

(ii) *The no-concept reply*: Another type-A response holds that the reason Mary lacks knowledge of what it is like to see red is simply that she fails to possess the relevant phenomenal concept. On this account, the conditional ' $PT \square Q$ ' is a priori, in that it is knowable a priori by anyone who possesses the concepts involved: it is just that Mary lacks one of the crucial concepts. If so, premise (1) is false, and the argument fails. This reply has not received much attention in the literature to date (although see Harman 1990, Hellie 2004, and Tye 2000 for suggestions in the vicinity), but in my view it is one of the more powerful replies available to a materialist.

Still, there are natural objections to this reply. While it is plausible that Mary lacks the phenomenal concepts involved in Q , it is less plausible that giving Mary this concept will close all relevant epistemic gaps. For a start, an opponent might appeal to the conceivability of zombies and inverted spectra: these suggest that even if one possess the concepts involved in P , T , and Q , there is no contradiction in the hypothesis that $PT \& \sim Q$, so that ' $PT \square Q$ ' is not a priori. One can also argue that once Mary has the relevant phenomenal concept, she will not automatically know whether or not *other* organisms (bats or Martians, say) are having experiences of the relevant sort, even given a complete physical description of them.

Similarly, one can mount a version of the knowledge argument using weaker phenomenal concepts that Mary possesses inside the black-and-white room. One such concept is that of phenomenal indistinguishability (Lahav 1994). Mary might possess all the physical facts, and possess the concept of phenomenal indistinguishability, while still being unable to know whether two subjects are having phenomenally indistinguishable experiences. If so, an argument analogous to the knowledge argument will go through.

Finally, one can make the case that even once Mary has emerged from the black-and-white room and has the relevant phenomenal concept, she cannot deduce the relevant phenomenal knowledge (e.g. that she is having an experience with such-and-such character) by a priori reasoning from PT alone. Rather, she must crucially rely on introspection. Introspection yields a posteriori knowledge, justified by experience rather than by reason alone. If this is correct, then ' $PT \square Q$ ' is not a priori.

(iii) *The indexical reply*: According to this analysis (Bigelow and Pargetter 1990, Ismael 1999, Perry 2001), Mary's new knowledge is likened to indexical knowledge. Proponents of this position will deny premise (1) above. They accept that ' $PT \square Q$ ' is a posteriori (and hence are phenomenal realists), but they deny that ' $PTI \square Q$ ' is a posteriori: Q is itself indexical knowledge, so if PTI contains full indexical knowledge it will entail Q . Mere indexical knowledge of her identity and the current time will obviously not help Mary to know what it is like to see red, but a proponent might appeal to further aspects of I . In particular, it is not implausible that I needs to build in further indexical information, identifying the referent of indexical phenomenal concepts such as E . Even so, this does not help the physicalist, for reasons discussed earlier. As we have seen already, Mary's central new knowledge does not involve any indexical element, so indexical information about the referent of E is distinct from Mary's new knowledge.

(One might think that even if Mary's new what-it-is-like knowledge is non-indexical, an indexical claim contained within I might help her derive this knowledge. In particular, a claim of the form *E is such-and-such* might be thought to help, if the right-hand side has the appropriate form: for example, if the indexical

claim is $E=R$, it will enable what-it-is-like knowledge. However, if E is *such-and-such* is to be built into PTI , then the right-hand-side must be such that *such-and-such is instantiated* is itself deducible from PT . (The indexical claims in I simply locate indexical referents on the objective map given by PT ; see Chalmers and Jackson 2001 for more here.) So given that non-indexical what-it-is-like knowledge of the form R is *instantiated* is not deducible from PT , then $E=R$ cannot be built into PTI . And it is clear that claims of the form $E=X$, where X is *instantiated* is deducible from PT (e.g. because X is a physical-functional concept) does not help Mary to deduce the relevant what-it-is-like knowledge. So if PT does not imply Q , neither does PTI .)

(iv) *The incomplete-physical-knowledge reply*: It is occasionally held that the knowledge argument fails as Mary did not have complete physical knowledge inside the black-and-white room. In effect, this reply will question premise (1) of the arguments I have given: rather than showing that ' $P \square Q$ ' is not a priori, the Mary case shows only that ' $P^* \square Q$ ' is not a priori, where P^* encompasses the subset of physical truths that Mary knows inside the room.

There are different versions of this reply, depending on how the proponent understands the notion of "physical" and the relevant incompleteness. Some proponents (e.g. Horgan 1984) are in effect using "physical" *broadly*, so that the physical truths encompass high-level truths that are necessitated by microphysical truths. Understood this way, it is nontrivial that knowing all the truths about physics and chemistry (and so on) suffices to know all the physical truths. A physicalist may hold that phenomenal truths themselves are broad physical truths, of which Mary is ignorant. On this view, it begs the question to assert the premise that Mary knows all the physical truths.

To avoid this issue, I think it is best to understand "physical" *narrowly*, as I do above. On this understanding, the physical is understood as the *microphysical* (or if one likes, the microphysical and the chemical, or some other specified domain). In this sense, it is less arguable that Mary knows all the physical truths: certainly we can stipulate that she knows all the truths in the language of microphysical theory. Of course in this sense, even high-level biological facts and the like will be nonphysical facts, so the existence of nonphysical facts is not enough to defeat physicalism. But the stronger claim that there are facts not *necessitated* by the microphysical facts is enough to defeat physicalism. And this stronger claim is delivered by the argument above.

(The knowledge argument is sometimes formalized as a straightforward inference from the fact that Mary knows all the physical facts but does not know all the phenomenal facts to the conclusion that phenomenal facts are not physical facts. I think this formulation does not provide a compelling argument against physicalism, for the reasons just stated. If "physical" is understood broadly, the claim that Mary knows all the physical facts is question-begging; if "physical" is understood narrowly, the conclusion that there are nonphysical facts is compatible with physicalism. For this reason, I think it is much better to formalize the knowledge argument in terms of deducibility and necessitation.)

Another version of this view (e.g. Stoljar 2001) holds that even if "physical" is restricted to the microphysical, Mary may nevertheless lack physical knowledge. On this view, Mary knows all truths in the language of microphysical theory, but there is more to microphysics than microphysical theory. In particular, microphysical theory gives Mary knowledge of the structural and relational properties of microphysical entities, but not of their intrinsic properties. And these intrinsic properties may be crucial to the necessitation of consciousness. Clearly, this view leads directly to the position that I earlier called "panprotopsychoism". As before, one can argue about whether this counts as a version of physicalism, but in any case I think it is a view that the knowledge argument leaves open.

(v) *The old-fact/new-way reply*: According to the most popular response to the knowledge argument, Mary gains knowledge of a fact she already knew, under a different mode of presentation (e.g. Horgan 1984; Loar 1997; Tye 2000). In the standard version, this is held to be analogous to someone who knew Hesperus is a planet and who learns that Phosphorus is a planet, or someone who knew that Superman can fly and who learns that Clark Kent can fly. Each pair of items of knowledge arguably involves a single fact (about a single property instantiated by a single individual), under distinct modes of presentation. Proponents of

this response hold that analogously, Mary's new phenomenal knowledge is knowledge of a fact she knew already (about the instantiation of a physical property), under a different mode of presentation. Mary has distinct physical and phenomenal concepts with a common referent, just as with *Hesperus* and *Phosphorus*, or *Superman* and *Clark Kent*.

The two-dimensional argument I have given above is formulated in such a way that it already takes this sort of response into account. In particular, all of the standard cases of new knowledge (involving *Hesperus*, *Superman*, and so on) seem to be compatible with premise 2: the claim that a posteriori statements have metaphysically contingent epistemic intensions. The a posteriori identity *Hesperus is Phosphorus* has a necessary subjunctive intensions, but it has a contingent epistemic intensions: the two concepts involved have distinct epistemic intensions across possible worlds, and the epistemic intension of the identity is correspondingly false at some world (e.g., a world where the evening star is distinct from the morning star). The same goes for all the other standard a posteriori identities: none give grounds for questioning premise 2, or any other premise of the argument above. So as it stands, this response does nothing to cast doubt on the knowledge argument as formulated here.

It is also possible to make this sort of point (somewhat less precisely) in ways that do not invoke the two-dimensional framework. For example, one can put the basic point by saying that where two non-indexical concepts *a* and *b* involve distinct modes of presentation of a common referent (i.e. when $a=b$ is not a priori), the distinct modes of presentation are associated in some fashion with distinct properties (connected only contingently) of the referent. When one gains new knowledge equating the referents of the two concepts, one gains knowledge of new contingent facts connecting the modes of presentation. In particular, one gains knowledge that a single individual has both associated properties. For example, when one learns that *Hesperus* is *Phosphorus*, one learns the new fact that the brightest object visible in the evening is also visible in the morning. When one learns that *Superman* is *Clark Kent*, one learns the new fact that the man with the cape works at the *Daily Planet*. When one learns that water is H_2O , one learns that the liquid in one's pool has a certain molecular structure. And so on. The thesis does not apply to indexical concepts (e.g. an objectively omniscient but indexically ignorant subject who learns *I am X*), but we have seen that indexicality is irrelevant in the case of phenomenal concepts.

The distinct-property thesis is closely related to the 2D Thesis stated above. The 2D Thesis entails that for an a posteriori identity $a = b$, *a* and *b* have distinct epistemic intensions across possible worlds. Here, epistemic intensions play much the same role as the associated properties invoked in the distinct-property thesis. The two-dimensional formulation has the advantages that it avoids the imprecise notion of "association", and that it accommodates indexical cases. (A concept such as *I* need not be associated with a property distinct from those associated with cognitively distinct coreferential "objective" concepts, but it does have a distinct epistemic intension.) On the other hand, the distinct-property thesis has the advantage that it requires less semantic theory for its formulation.

In effect the distinct-property thesis allows us to find a new fact associated with any case of new knowledge, so that if Mary gains new knowledge of an old fact, she also gains knowledge of a new fact. One could also run the argument directly, by invoking something like the following thesis:

New Fact Thesis: In non-indexical cases, whenever one gains new knowledge of an old fact, one simultaneously gains knowledge of some new fact.

This principle (versions of which are put forward by Lockwood 1989, pp. 136-7, Chalmers 1996, pp. 141-2, and Thau 2002, p. 127) is closely related to the 2D Thesis and the distinct-property thesis, but it requires neither two-dimensionalism nor the notion of association between concepts and properties. It can be endorsed even by those with very different views about reference and mental content. An examination of cases, of the sort given above, suggests that it is independently plausible. In all these cases, when one gains knowledge of an old fact under a new mode of presentation, one gains knowledge of a new fact connecting the two modes of presentation.

Applying this principle to Mary: if Mary acquires new knowledge of an old fact, she will also acquire knowledge of a new fact. In effect, the New Fact Thesis allows us to move from the epistemic claim that Mary gains new factual knowledge (of either an old or a new fact) to the ontological claim that Mary gains knowledge of a new fact. If so, if Mary already knew all the physical facts, then the physical facts do not exhaust all the facts.

Strictly speaking, this conclusion is compatible with physicalism, at least if we understand physical facts as microphysical facts. But to obtain a stronger conclusion, we need only refine the principle slightly.

New Fact Thesis (modified): In non-indexical cases, whenever one gains new knowledge (not deducible by a priori reasoning from previous knowledge) of an old fact, one simultaneously gains knowledge (or becomes in a position to gain knowledge) of some fact not necessitated by previously known facts.

Once again, this thesis is plausible in all the standard cases. Applying it to Mary: given that Mary gains new knowledge, not derivable from her previous knowledge, it follows that she gains knowledge (or becomes in a position to gain knowledge) of some fact not necessitated by previously known facts. Given that she previously knew all the microphysical facts, it follows that there are facts not necessitated by the microphysical facts, so that physicalism is false.

(vi) *Loar's reply.* For an old-fact/new-way analysis to have any hope of succeeding, it must treat Mary's new knowledge as disanalogous to standard cases of coreference, and in particular it must give reason to think that premise 2 fails in this case, as do the related principles discussed above. This is an uphill battle, since it is plausible that these principles hold in all familiar cases. In particular, all familiar a posteriori identities appear to involve distinct epistemic intensions over centered worlds.

A sophisticated attempt in this direction is made by Loar (1997). Loar isolates the "semantic premise" of the knowledge argument:

Semantic premise: A statement of property identity that links conceptually independent concepts is true only if at least one concepts picks out the property it refers to by connoting a contingent property of that property. (Loar 1997, p. 600)

Here, two concepts a and b are conceptually independent when an identity judgment $a=b$ is a posteriori. A concept connotes a property when it uses that property to pick out its referent. A concept's connoted property is akin to a property associated with the concept's mode of presentation in the sense discussed above. If we make the translation, Loar's semantic premise is closely akin to the distinct-property thesis discussed above, according to which the concepts involved in an a posteriori identity are associated with contingently coextensive properties. Strictly speaking the distinct-property thesis is weaker than Loar's semantic premise, as distinct properties can be contingently coextensive even when each is a non-contingent property of its bearer (e.g. the properties of containing hydrogen atoms and containing H₂O). There are corresponding exceptions to Loar's semantic premise (as Tye 1997 has pointed out), but (as White 2002 points out) the weaker premise that the concepts in such an identity connote contingently coextensive properties is enough to make the knowledge argument work.

(Like the distinct-property thesis, Loar's semantic premise is closely related to the 2D Thesis, although like the distinct-property thesis, it has exceptions in the case of indexicals.)

Loar allows that the semantic premise applies to standard a posteriori identities, but denies that it holds in the physical-phenomenal case. He suggests (p. 602) that this can be explained by the fact that phenomenal concepts are recognitional concepts that pick out physical properties without a contingent associated mode of presentation. Here, a recognitional concept is a type-demonstrative ("one of *that* kind"). Loar holds that the fact that phenomenal concepts are recognitional concepts explains the aposteriority of a phenomenal-physical identity: identities involving recognitional and theoretical concepts are always a posteriori.

Most such identities satisfy the semantic premise, but Loar holds that an exception in this case is explained by the fact that phenomenal concepts lack a contingent mode of presentation. Loar's idea of a concept that lacks a contingent mode of presentation corresponds approximately, on the two-dimensional account, to what I called semantically neutral concept above: roughly, a concept with the same epistemic and subjunctive intensions across metaphysically possible worlds (Loar 1999 makes the connection explicitly). Here we can call these *neutral* concepts for short.

Loar's key claims are, in effect: (i) phenomenal concepts are neutral and recognitional; (ii) if phenomenal concepts are recognitional, they will be involved in a posteriori identities with theoretical concepts; (iii) if phenomenal concepts are neutral, these a posteriori identities will not involve contingent modes of presentation. From these claims it follows that the semantic premise is false (as is premise 2), so that the epistemic gap will be compatible with physicalism.

I think there is good reason to reject this account. First, as before, there is good reason to believe that the phenomenal concepts crucial to Mary's argument are not type-demonstratives. Mary's demonstrative phenomenal concepts are type-demonstratives, but her pure phenomenal concepts are quite distinct from these, and it is pure phenomenal concepts that are crucial to the argument. So an account based on an appeal to type-demonstratives cannot succeed here. Loar might respond that he is using "type-demonstrative" more broadly than I am, so that pure phenomenal concepts still count as type-demonstratives. Or he could drop the appeal to demonstrative concepts, and hold that recognitional concepts need not be demonstrative. But a deeper problem remains.

Let us say that a concept is *opaque* when it can corefer with a conceptually independent neutral theoretical concept. In effect, Loar argues that pure phenomenal concepts are both neutral and opaque: they are themselves neutral, but they can corefer with conceptually independent neutral theoretical concepts. To make discussion easier, I will play along with Loar's view that the relevant theoretical concepts are themselves neutral. (If they are non-neutral, this complicates matters slightly, but in the end it opens up room only for panprotopsychism.)

If pure phenomenal concepts were both neutral and opaque, the semantic premise (and premise 2) would be false, and the anti-materialist argument would fail. But no reason has been given to believe that any concept can be both neutral and opaque. Every other clear case of an opaque concept (demonstrative concepts, many natural-kind concepts, opaque recognitional concepts) is non-neutral, and their non-neutrality underlies and explains their opacity. By contrast, every other clear case of a neutral concept (certain descriptive concepts, categorical concepts) appears to be non-opaque.

Anticipating a response along these lines (p. 602), Loar says (in effect) that antiphenomenalists hold that phenomenal concepts are neutral, and that the physicalist is entitled to the same claim. But the plausibility of the claim that phenomenal concepts are neutral does nothing to resolve the conflict between that claim and the claim that phenomenal concepts are opaque. Insofar as we have reason to believe that phenomenal concepts are neutral, we have reason to believe that they are not opaque, and so we have reason to believe that physicalism is false (or that panprotopsychism is true).

Even if there is a sense in which phenomenal concepts are recognitional, recognitionality alone does nothing to support opacity. The other recognitional concepts that Loar appeals to in order to support the claim of opacity are *non-neutral* recognitional concepts, and it is their non-neutrality that grounds their opacity. (That is, it is because these concepts present their referent under a contingent mode of presentation that they can corefer with conceptually independent neutral concepts.) So if phenomenal concepts are neutral, these analogies give no reason to accept that phenomenal concepts are opaque. We have good reason to believe that all neutral concepts are non-opaque, and no reason has been given to grant an exception.

To sum up: Loar's claims of (i) conceptually independent coreference (opacity) and (ii) lack of contingent mode of presentation (neutrality) stand in strong tension with each other, a tension that Loar's account does

nothing to remove. Once we accept the second claim, any support for the first claim is undercut. So there is no reason to believe that the relevant phenomenal concepts have the features that Loar suggests, and there is good reason to deny this claim. I conclude that Loar's attempt to reconcile the distinctive epistemic behavior of phenomenal concepts with physicalism fails.

(Further discussion of Loar's account is given in Chalmers 1999. I have not stressed modal considerations here, but I argue there that Loar's account requires "strong necessities" and a sort of modal dualism that is quite problematic in its own right.)

(vii) *Other analyses.* The materialist might hope that some other account of phenomenal concepts can be given, such that their distinctive epistemic behavior can be reconciled with the claim that they refer to physical properties. I am skeptical that any such account can be given. I think the only remote chance is to attempt to deny premise 2 (and related claims about properties associated with modes of presentation), but as I have argued elsewhere (Chalmers 1999), there is reason to believe that this premise is a deep (nontrivial) conceptual truth. On my view, all other accounts that attempt to deny this premise fall prey to problems that are akin to the problem of Loar's account.

Nevertheless, there is room for fruitful further debate on this topic. And whatever the consequences for the truth of materialism, a deeper understanding of phenomenal concepts is likely to have deep consequences for our understanding of consciousness.

References

- Austin, D.F. 1990. *What's the Meaning of "This"?* Ithaca, NY: Cornell University Press.
- Bigelow, J. and Pargetter, R. 1990. Acquaintance with qualia. *Theoria*.
- Chalmers, D.J. 1996. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford University Press.
- Chalmers, D.J. 1999. Materialism and the metaphysics of modality. *Philosophy and Phenomenological Research*.
- Chalmers, D.J. 2002a. Consciousness and its place in nature. In (S. Stich & F. Warfield, eds) *The Blackwell Guide to the Philosophy of Mind*. Blackwell. [consc.net/papers/nature.html]
- Chalmers, D.J. 2002b. The components of content. In (D. Chalmers, ed) *The Philosophy of Mind: Classical and Contemporary Readings*. Oxford University Press. [consc.net/papers/content.html]
- Chalmers, D.J. 2002c. The content and epistemology of phenomenal belief. In (Q. Smith & A. Jokic, eds) *Consciousness: New Philosophical Essays*. Oxford University Press.
- Chalmers, D.J. (forthcoming). The nature of epistemic space. [consc.net/papers/espace.html]
- Chisholm, R. 1957. *Perceiving: A Philosophical Study*.
- Francescotti, R.M. 1994. Qualitative beliefs, wide content, and wide behavior. *Nous* 28:396-404.
- Hawthorne, J. 2002. Advice to physicalists. *Philosophical Studies* 109:17-52.
- Hellie, B. 2004. Inexpressible truths and the allure of the knowledge argument. In (P. Ludlow, Y. Nagasawa, & D. Stoljar, eds) *There's Something About Mary*. MIT Press.

- Horgan, T. 1984. Jackson on physical information and qualia. *Philosophical Quarterly* 34:147-83.
- Ismael, J. 1999. Science and the phenomenal. *Philosophy of Science*.
- Jackson, F. 1982. Epiphenomenal qualia. *Philosophical Quarterly* 32:127-136.
- Kaplan, D. 1989. Demonstratives. In (J. Almog, J. Perry, and H. Wettstein, ed.) *Themes from Kaplan*. New York: Oxford University Press.
- Lahav, R. 1994. A new challenge for the physicalist: Phenomenal indistinguishability. *Philosophia* 24:77-103.
- Loar, B. 1997. Phenomenal states (second version). In (N. Block, O. Flanagan, & G. Güzeldere, eds) *The Nature of Consciousness*. MIT Press.
- Loar, B. 1999. David Chalmers' *The Conscious Mind*. *Philosophy and Phenomenological Research* 59:464-71.
- Lockwood, M. 1989. *Mind, Brain, and the Quantum*. Blackwell.
- Nida-Rümelin, M. 1995. What Mary couldn't know: Belief about phenomenal states. In (T. Metzinger, ed) *Conscious Experience*. Ferdinand Schöningh.
- Perry, J. 2001. *Knowledge, Possibility, and Consciousness*. MIT Press.
- Stoljar, D. 2001. Two conceptions of the physical. *Philosophy and Phenomenological Research* 62:253-81.
- Thau, M. 2002. *Consciousness and Cognition*. Oxford University Press.
- Tye, M. 1997. Qualia. *Stanford Encyclopedia of Philosophy* (fall 1997 edition).
[plato.stanford.edu/archives/fall1997/entries/qualia]
- Tye, M. 2000. Knowing what it is like: The ability hypothesis and the knowledge argument. In *Consciousness, Color, and Content*. MIT Press.
- White, S. 2002. Why the property dualism argument won't go away. *Journal of Philosophy*.

Appendix

What follows is a brief and simplified introduction to the two-dimensional semantic framework as I understand it. See also Chalmers (2002b; forthcoming).

Let us say that it is epistemically possible in the broad sense that S if the hypothesis that S is not ruled out a priori. Then there will be a wide space of epistemic possible hypotheses (in the broad sense; I omit the qualifier in what follows). Some of these will conflict with each other; some of them will be compatible with each other; and some will subsume each other. We have a systematic way of describing epistemic possibilities that differs from our way of describing subjunctive counterfactual possibilities. It is this sort of epistemic description that is captured by the first dimension of the two-dimensional framework.

It is epistemically possible that water is not H₂O, in the broad sense that this is not ruled out a priori. And there are many specific versions of this epistemic possibility: intuitively, specific ways our world could turn out such that if they turn out that way, it will turn out that water is not H₂O. Take the XYZ-world, one

containing superficially identical XYZ in place of H₂O. It is epistemically possible that our world is the XYZ-world. When we consider this epistemic possibility — that is, when we consider the hypothesis that *our* world contains XYZ in the oceans, and so on — then this epistemic possibility can be seen as an instance of the epistemic possibility that water is not H₂O. We can rationally say "if our world turns out to have XYZ in the oceans (etc.), it will turn out that water is not H₂O". The hypothesis that the XYZ-world is actual rationally entails the belief that water is not H₂O, and is rationally inconsistent with the belief that water is H₂O.

Here, as with subjunctive counterfactual evaluation, we are considering and describing a world, but we are considering and describing it in a different way. In the epistemic case, we consider a world *as actual*: that is, we consider the hypothesis that our world is that world. In the subjunctive case, we consider a world *as counterfactual*: that is, we consider it as a way things might have been, but (probably) are not. These two modes of consideration of a world yield two ways in which a world might be seen to make a sentence or a belief true. When the XYZ-world is considered as actual, it makes 'water is XYZ' true; when it is considered as counterfactual, it does not.

In considering a world as actual, we ask ourselves: what if the actual world is really that way? In the broad sense, it is *epistemically* possible that Hesperus is not Phosphorus. This is mirrored by the fact that there are specific epistemic possibilities (not ruled out a priori) in which the heavenly bodies visible in the morning and evening are distinct; and upon consideration, such epistemic possibilities are revealed as instances of the epistemic possibility that Hesperus is not Phosphorus.

When we consider worlds as counterfactual, we consider and evaluate them in the way that we consider and evaluate subjunctive counterfactual possibilities. That is, we acknowledge that the character of the actual world is fixed, and say to ourselves: what if the world *had been* such-and-such a way? When we consider the counterfactual hypothesis that the morning star might have been distinct from the evening star, we conclude not that Hesperus would not have been Phosphorus, but rather that at least one of the objects is distinct from both Hesperus and Phosphorus (at least if we take for granted the actual-world knowledge that Hesperus is Phosphorus, and if we accept Kripke's intuitions).

Given a statement *S* and a world *W*, the *epistemic intension* of *S* returns the truth-value of *S* in *W* considered as actual. (Test: if *W* actually obtains, is *S* the case?) The *subjunctive intension* of *S* returns the truth-value of *S* in *W* considered as counterfactual. (Test: if *W* had obtained, would *S* have been the case?) We can then say that *S* is *primarily possible* (or 1-possible) if its epistemic intension is true in some world (i.e. if it is true in some world considered as actual), and that *S* is *secondarily possible* (or 2-possible) if its subjunctive intension is true in some world (i.e. if it is true in some world considered as counterfactual). Primary and secondary necessity can be defined analogously.

For a world to be considered as actual, it must be a *centered* world — a world marked with a specified individual and time — as an epistemic possibility is not complete until one's "viewpoint" is specified. So a epistemic intension should be seen as a function from centered world to truth-values. For example, the epistemic intension of 'I' picks out the individual at the center of a centered world; and the epistemic intension of 'water' picks out, very roughly, the clear drinkable (etc.) liquid in the vicinity of the center. No such marking of a center is required for considering a world as counterfactual, or for evaluating subjunctive intensions.

Epistemic and subjunctive intensions can be associated with statements in language, as above, and equally with singular terms and property terms. The intension of a statement will be a function from worlds to truth-values; the intension of a term will be a function from worlds to individuals or properties within those worlds. (In some cases, intensions are best associated with linguistic tokens rather than types.)

Epistemic intensions can also be associated in much the same way with the (token) concepts and thoughts of a thinker, all of which can be used to describe and evaluate epistemic possibilities as well as subjunctive counterfactual possibilities. In "The Components of Content" I argue that the epistemic intension of a

concept or a thought can be seen as its "epistemic content" (a sort of internal, cognitive content), and that the subjunctive intension captures much of what is often called "wide content".

A crucial property of epistemic content is that it reflects the rational relations between thoughts. In particular, if a belief A entails a belief B by a priori reasoning, then it will be epistemically impossible (in the broad sense) for A to be true without B being true, so the epistemic intension of A entails the epistemic intension of B . Further, if an identity $a=b$ is a posteriori for a subject, then it is epistemically possible for the subject that the identity is false, and there will be an epistemic possibility in which the referents of the two concepts involved differ, so the subject's concepts a and b will have distinct epistemic intensions. So epistemic intensions behave something like Fregean senses, individuating concepts according to cognitive significance at least up to the level of a priori equivalence. **01175731**