The Philosophical Review, Vol. 108, No. 1 (January 1999)

Conceptual Analysis, Dualism, and the Explanatory Gap

Ned Block and Robert Stalnaker

1. Introduction

One point of view on consciousness is constituted by two claims:

> **The explanatory gap.** Consciousness is a mystery. No one has ever given an account, even a highly speculative, hypothetical, and incomplete account of how a physical thing could have phenomenal states (Nagel 1974, Levine 1983). Suppose that consciousness is identical to a property of the brain—say, activity in the pyramidal cells of layer 5 of the cortex involving reverberatory circuits from cortical layer 6 to the thalamus and back to layers 4 and 6—as Crick and Koch have suggested for visual consciousness (Crick 1994). Still, that identity itself calls out for explanation! Proponents of an explanatory gap disagree about whether the gap is permanent. Some (e.g., Nagel 1974) say that we are like the scientifically naive person who is told that matter = energy, but does not have the concepts required to make sense of the idea. If we can acquire these concepts, the gap is closable. Others say the gap is unclosable because of our cognitive limitations (McGinn 1991). Still others say that the gap is a consequence of the fundamental nature of consciousness.
>
> **No conceptual analysis.** Some concepts are analyzable functionally, or in terms of the concepts of physics. Perhaps even some mental concepts can be given functional or physical analyses. But consciousness is not one of these analyzable concepts. Further, this unanalyzability is no accident: any putative functional or physical analysis would leave out the fundamental na-

We are grateful to the following persons for their helpful comments on previous drafts of this paper: Alex Byrne, David Chalmers, Frank Jackson, Joe Levine, Barry Loewer, and the editors and referees for the *Philosophical Review.*

ture of consciousness. Because there is no conceptual analysis of consciousness in physical or functional terms, there is no contradiction in the notion of a zombie that is a functional duplicate—or even a microphysical duplicate—of one of us, but that has no consciousness at all.

Our main concern is with the relation between these two claims: specifically, with the relation between the claim that there is an unclosable explanatory gap as a result of the fundamental nature of consciousness and the claim that there is no conceptual analysis of consciousness in functional or physical terms. It should be uncontroversial that the first entails the second, for if the concept of consciousness were functionally analyzable, we could close the explanatory gap by showing how that functional role could be physically implemented. What is more controversial is whether the claim that there is no conceptual analysis of consciousness entails that there is an explanatory gap that can never be closed. We will call this position—that conceptual analysis is *necessary* to close the explanatory gap—the *conceptual analysis thesis*. It is shared by a number of philosophers whose overall responses to the problem of consciousness is otherwise quite different. For example, Joseph Levine differs on the metaphysical consequences of the explanatory gap from Frank Jackson and David Chalmers. Levine argues that the gap is an epistemological one that is compatible with the thesis that facts about consciousness supervene on the physical facts, while Jackson and Chalmers argue that the fact that there can be no conceptual analysis of consciousness supports metaphysical dualism: consciousness is neither identical with nor supervenient on the physical. We will be criticizing the conceptual analysis thesis and the further claim that the lack of a conceptual analysis of consciousness entails dualism. We will be paying more attention to the views shared by the proponents of these theses than to their differences.[1]

[1]Horgan 1984 (building on Lewis 1983) argues that a Laplacean demon could figure out all the facts from the microphysical facts and meaning constraints. Versions of the view that facts that don't involve consciousness follow a priori from physics appear in Jackson 1993, 1994, 1995; Levine 1983, 1993; and Chalmers 1996. (Important precursors are to be found in Lewis 1966 and Nagel 1974.) Jackson 1994 is a variant of Jackson 1993. Jackson 1995 is a brief and lucid summary of the same points in the context of Jackson's "Mary" argument against physicalism. Two Ph. D. theses, Byrne 1993 and Chalmers 1993, cover much the same ground.

2. The Epistemic Version

The arguments for the existence of an explanatory gap between the mental and the physical standardly rely on thought experiments that purport to show that certain situations (for example, the existence of a mind just like mine without a body, or of a body in the same physical state as mine when I am feeling pain, but without anyone feeling pain) are possible. But Levine emphasizes that the metaphysical intuitions on which such conclusions rest are controversial. His strategy is to argue for an explanatory gap on more cautious assumptions—assumptions that are compatible with the truth of physicalism. The explanatory gap, he argues, is epistemological rather than metaphysical: it is a gap in our understanding of *how* the physical facts make the mental facts true, a gap that would not be closed even if we accepted the thesis that the mental facts *are* made true by the physical facts. Levine's argument makes use of the same kinds of thought experiments, but it takes them to be about what is merely conceivable or imaginable rather than about what is metaphysically possible. He claims that conceivability arguments, even if insufficient to establish a metaphysical conclusion, can still show that a certain kind of explanation of mental phenomena in terms of the physical is unavailable.

After sketching Levine's argument, we will look closely at just what is meant by conceivability, at the relation between conceivability and possibility, and at the relation between what is conceivable and what is compatible with conceptual truths. We will contrast two ways of understanding conceivability. On one, intuitions about what is conceivable are at least as problematic and controversial as intuitions about what is metaphysically possible. On the other, we will grant that the case has been made for a conceivability gap between the mental and the physical, but argue that conceivability in this sense is insufficient for an explanatory gap.

Both Levine and Jackson recognize and are responding to the distinction brought out in Kripke 1972 between what is metaphysically necessary and what is a priori. Some necessary truths, such as that water = H_2O, are not a priori truths, and so despite their necessity, their truth cannot be established by analysis of the relevant concepts. We can imagine discovering that water is something other than H_2O, and so in a sense it is conceivable, even though impossible, that water is not H_2O. The existence of necessary a

posteriori truths shows that there is no simple and direct path from the conceptual independence of consciousness and the physical to their metaphysical independence. As Kripke emphasized, a posteriori necessities give rise to an illusion of contingency that needs to be explained away, but as Levine emphasizes, the fact that the appearance of contingency is sometimes an illusion shows the fragility of intuitions about metaphysical possibilities. Levine in fact rejects the assumption that our intuitions give us access to metaphysical reality, arguing that "one's ideas can be as clear and distinct as you like, and nevertheless not correspond to what is in fact possible" (1993, 123). It is for this reason that he wants his argument to remain neutral on metaphysical questions, such as whether consciousness is in fact identical to pyramidal cell activity. But it is essential to his argument that it bring out an asymmetry between the water/H_2O case and the consciousness/pyramidal cell activity case, since the point is to show that physical theory cannot explain the phenomena of consciousness in the way that it can explain the behavior of water.

Levine's argument rests on an account of how the explanatory gap is closed in the paradigm cases of satisfactory explanation. Consider the question, Why does water boil when it is heated? Here is a rough sketch of an answer: The molecular kinetic energy of the H_2O molecules increases, causing more and more molecules to escape from the liquid, forming bubbles within the liquid. The average momentum of these molecules is a kind of pressure ("vapor pressure"), which increases as the temperature increases at a rate dependent on the strength of the bonds between molecules. When the vapor pressure = the atmospheric pressure, bonds break throughout the water, causing H_2O vapor to escape from the surface and to form globules that bubble up.

According to Levine, this is a sketch of an adequate explanation of why water boils under certain conditions because it shows how it can be deduced from microphysics and chemistry that water boils in certain conditions. But since the word 'boil' is not a term of the microphysical and chemical theories, how is this to be done? "The problem," Levine says, "is that chemical theory and folk theory don't have an identical vocabulary, so somewhere one is going to have to introduce bridge principles. . . . We need a definition of 'boiling' and 'freezing' that brings these terms into the proprietary vocabularies of the theories appealed to in the explanation" (1993,

131). With the help of such definitions we can deduce from chemistry and physics answers to questions stated in our ordinary folk vocabulary. "On this view, explanatory reduction is, in a way, a two-stage process. Stage 1 involves the (relatively? quasi?) *a priori* process of working the concept of the property to be reduced 'into shape' for reduction by identifying the causal role for which we are seeking the underlying mechanisms. Stage 2 involves the empirical work of discovering just what those underlying mechanisms are" (1993, 132).[2]

Levine of course recognizes that many of the required bridge principles connecting folk with scientific vocabulary (such as that water is H_2O, and that boiling is the particular microphysical process that it is) will not be analytic definitions: very often, what is needed are the notorious necessary a posteriori truths. But Levine's original point was that metaphysical necessity alone would not suffice to close an explanatory gap; his claim is that for explanation, we need a *deduction* of the phenomenon to be explained from a lower level explanatory science, with the help of bridge principles that are provided by a priori conceptual analysis—the kind of conceptual analysis that cannot be given for consciousness. We need to show, not just that it is *impossible* (given our scientific theory) for water not to boil in the relevant circumstances, but that it is *inconceivable* that it not boil in those circumstances.

It is clear enough that the kind of explanation sketched above removes any mystery about why water boils, but what reason is there to think that conceptual analyses of 'water' and 'boiling' are implicit in the story? It may be necessary that boiling is the particular physical process with which the explanation identifies it, but if so, it is clearly necessary a posteriori. The stuff they call 'water' on Twin Earth does something Twin Earthians call 'boiling', but XYZ does not therefore do what *we* call 'boiling'. Levine grants that we cannot reason a priori from the existence of boiling water

[2]As is clear from this quotation, and also from other passages in this paper, Levine has some reservations about the conceptual analysis thesis. If one concedes that the analyses that result from the process of working ordinary concepts into shape for reduction are only relatively or quasi a priori, then (depending on how the qualifications are developed) it may be more plausible to assume that terms like 'water' and 'boiling' have analyses. But it may at the same time become more plausible to say that consciousness has an analysis of this kind in physical terms.

to the existence of H_2O undergoing the particular physical process that constitutes boiling, but he argues that we *can* reason a priori in the other direction—from microphysical theory and fact to the presence of water and the realization of the property of boiling—and that this reasoning reveals a contrast with the case of conscious states and brain processes.

Levine does not provide an actual analysis of boiling that would support this claim, but instead appeals to intuitions about conceivability. "While it is conceivable that something other than H_2O should manifest the superficial macro properties of water, as Kripke suggests, it is not conceivable, I contend, that H_2O should fail to manifest these properties (assuming of course that we keep the rest of chemistry constant)" (1993, 128). Nothing (holding our chemistry and physics constant) could conceivably be H_2O and not, for example, boil under the appropriate conditions. Moreover, nothing could conceivably instantiate the molecular motions that are actually characteristic of water boiling (vapor molecules bubbling off the surface of liquid H_2O) without being *boiling*. However, according to Levine, one cannot say the analogous thing about conscious states and their neural correlates. Even if pain turns out to be perfectly correlated with pyramidal cell activity, and even if we decide that pain *is* (necessarily) pyramidal cell activity, it will remain possible (Levine contends) to conceive of pyramidal cell activity without pain.

But such claims about conceivability seem at least as fragile and fallible as intuitions about what is metaphysically possible. What exactly does it mean to say that we can conceive of something even if it may in fact be impossible? The intuition that we can make sense of this is fueled by Kripke's cases of necessary a posteriori truths: it is conceivable (even if impossible) that water should turn out not to be H_2O, or that Queen Elizabeth II should be the daughter of Bess and Harry Truman (Kripke 1972). But as Kripke's discussion makes clear, these are cases of misdescribed possibilities. What we imagine, or conceive of, in these cases are genuine metaphysical possibilities—they are not just the possibilities that water is something other than H_2O, or that Elizabeth herself has different parents. What lies behind Kripke's cases is the fact that the meaning and reference of our terms depend on empirical facts, facts that we might be ignorant or mistaken about. What we conceive of when we conceive of a possibility we describe as one in

which water is something other than H_2O is a genuine possibility, a possible world in which as speakers *there* use the term 'water' it refers to something other than H_2O—and so something other than water. (This is the first of our two notions of conceivability.) Now if, in describing a possible world, we stipulate that all the facts on which the meaning and reference of certain terms depend are the same as they are in the actual world, then the possibility we are describing will of course be one in which those terms, as speakers use them *there,* refer to the same things they refer to when we use them. (We shall be making use of this point later.) So if we hold physics and chemistry, and the relevant particular facts fixed, we can be sure we have a possible situation in which the expression 'water is boiling' (as used by us, or by them) expresses a truth. Perhaps this fact shows a sense in which it is inconceivable that (holding physics and chemistry fixed) water should not boil in the relevant circumstances, but this has nothing to do with conceptual analysis, and it will not show that there is any asymmetry between the H_2O/boiling case and the pain/pyramidal cell case. Grant, for the moment, the metaphysical thesis that pain is (necessarily) identical to pca. Now consider a possible world in which the relevant physical theories and circumstances are held fixed (that is, are stipulated to be the same as in the actual world). 'Pain', as we use the term, obviously applies to pca in this possible world, but what about 'pain', as used by the people in this counterfactual world? Can we consistently say about such a possible world that the people in it, who are physically just like us, refer with 'pain' to something other than what we call 'pain?' It is not clear that we can, and if we cannot, then this way of thinking about conceivability without possibility does not show that pca without pain is conceivable, and so does not show any asymmetry.

Here is a different way of trying to make sense of conceivability without possibility, one that ties conceivability explicitly to what is compatible with concepts: One might say that *P* without *Q* is conceivable if it is not possible to deduce *Q* from *P,* using only logic and conceptual truths (such as truths that follow from conceptual analyses). Water in the bathtub without H_2O in the bathtub is conceivable because one cannot deduce from a correct conceptual analysis of water that it is H_2O. This is a purely negative conception of conceivability (and so the term is somewhat misleading). One might, for example, conclude that it is conceivable that Bill Clinton

is identical with Newt Gingrich simply on the ground that there are no analytic truths involving the proper names from which 'Bill Clinton is not Newt Gingrich' can be deduced.

Now on this account of conceivability, we think it will be right to say that even if pain is in fact pca, pca without pain is still conceivable, but on this account, Levine's argument about the water boiling example won't work. Let *C* be a complete description, in microphysical terms, of a situation in which water (H_2O) is boiling, and let *T* be a complete theory of physics. Can one deduce from *T,* supplemented with analytic definitions, that H_2O would boil in circumstances *C*? To see that one cannot, suppose that the deduction is taking place on Twin Earth. The stuff they call 'water' is XYZ, and the process they call 'boiling' is a process that superficially resembles boiling, but that involves a different physical process. Just as they would say (truly), "Water is XYZ, and not H_2O (and if there were H_2O, it wouldn't be water)," so they would say (truly), "If there were H_2O, and it were behaving like that, it wouldn't be boiling." They could hardly deduce 'H_2O would boil in circumstances C' if on their meaning of 'boil', H_2O can't boil at all. (We assume that boiling is a natural kind concept. If you don't agree, substitute some other process term that does express a natural kind concept.)

We don't really need a Twin Earth story to make our point. Consider a person on actual Earth, who does not know the story about how water boils—perhaps she doesn't even know that water is made up of molecules. One presents her with the theory *T,* and a description (in microphysical terms) of a water boiling situation. Can she then deduce that if *T* is true and a situation met conditions *C,* then the H_2O would be boiling? No, since for all she knows the actual situation is like the one on Twin Earth. Perhaps, if she were told, or could figure out, that the theory was actually true of the relevant stuff in her environment, she could then conclude (using her knowledge of the observable behavior of the things in her environment) that H_2O is water, and that the relevant microphysical description is a description of boiling, but the additional information is of course not a priori, and the inference from her experience would be inductive.

So we are not persuaded by Levine's argument, but we agree, nevertheless, with his sketch of the kind of explanation required for facts such as that water boils. All we reject is the a priori, purely

conceptual status attributed to the bridge principles connecting the ordinary description of the phenomena to be explained with its description in the language of science. What is actually deduced in such an explanation is a description wholly within the language of science of the phenomenon to be explained. For this to answer the original explanatory question, posed in so-called folk vocabulary, all we need to add is the claim that the phenomenon described in scientific language is the same ordinary phenomenon described in a different way. But if the closing of an explanatory gap does not require an a priori deduction of the folk description of the phenomena, then it has not been shown that unavailability of a conceptual analysis of consciousness need be an obstacle to the closing of the explanatory gap between consciousness and the physical.

3. The Metaphysical Version

Levine's argument tries to use the conceptual analysis thesis to *bypass* the metaphysical question about whether the mental supervenes on the physical. Frank Jackson and David Chalmers, in contrast, want to use the conceptual analysis thesis to *support* the claim that there is a metaphysical gap between mental and physical; it is part of an argument for a kind of dualism, for the conclusion that the facts about consciousness do not supervene on the physical facts. Like Levine, Jackson and Chalmers are concerned with the relation between ordinary pre-scientific terms—such as 'water', 'heat', and 'boiling'—and the terms of chemistry and microphysics, and with the role of the latter in the explanation of phenomena described in ordinary terms. Like Levine, they recognize that statements connecting the terms of the two kinds (such as that water is H_2O) are often both necessary and a posteriori, and so are not claims that can be justified on purely conceptual grounds. Thus, they agree with Levine that there is not in general any simple and direct inference from the conceptual independence of terms to the metaphysical independence of the properties expressed by the terms. It is agreed that the concepts expressed by 'water' and 'H_2O' are independent even though they name the same thing. But they argue that even with a posteriori necessities, the justification for the claim of necessity must be grounded in conceptual analysis. They argue that one can give conceptual analyses of terms like

'water' and 'heat', analyses that when conjoined with contingent empirical microphysical claims are sufficient to deduce the a posteriori necessities, and so to connect folk descriptions of phenomena with their scientific explanations. Since the kind of conceptual analysis that is available for terms like 'water' and 'heat' is not available for the terms expressing phenomenal concepts, we can conclude that physicalism is false: there is no metaphysically necessary connection between phenomenal consciousness and the physical, and no possibility of an explanation of phenomenal consciousness in physical terms.

The metaphysical thesis of physicalism, according to Jackson, can be defined as follows: any possible world that is a *minimal* physical duplicate of our world is a duplicate *simpliciter* of our world. A minimal physical duplicate of a world is one that is indiscernible from that world with respect to all physical objects, properties, and relations, and in addition contains nothing "extra," nothing that is not required in order to be a physical duplicate. The reason for the minimality requirement is this: physicalists may grant the metaphysical possibility of nonphysical stuff—ghostly ectoplasm for example—and so may grant the possibility of worlds that are physical duplicates of ours, but contain some nonphysical stuff as well. Since the thesis the physicalist wants to defend will be false in such possible worlds, the thesis must be formulated to as to exclude them.

Jackson claims that it follows from physicalism, understood this way, that "any psychological fact about our world is *entailed* by the physical nature of our world" (1993, 131; our emphasis). Entailment, as Jackson uses the term, is to be understood as a metaphysical rather than a logical relation: a set of premises entails a conclusion if and only if it is metaphysically necessary that if all propositions in the premise set are true, then the proposition expressed in the conclusion is true. Since some entailments are not a priori, Jackson recognizes that he needs additional argument to show that conceptual analysis of the mental in physical terms is required for a defense of physicalism. As Jackson says, "Conceptual analysis in the traditional sense . . . is constituted by *a priori* reflection on concepts and possible cases with an aim to elucidating connections between different ways of describing matters. Hence, it might be objected, if we allow that some entailments are *a posteriori,* to demonstrate an entailment is conspicuously not to demonstrate the importance of conceptual analysis" (1993, 136). The

main burden of Jackson's case is to provide this additional argument.

Jackson's central example will help to explain both the problem and his solution to it, and we will keep coming back to it. Consider the fact that the earth is covered 60 percent by water. We should be able to show, Jackson contends, that this fact is entailed a priori by the microphysical facts, which means that we should be able to deduce the statement that the earth is covered 60 percent by water from truths statable in the vocabulary of microphysics, together with truths knowable by a priori reflection. Suppose we can deduce "the earth is covered 60 percent by H_2O" from the microphysical facts. Then we can conclude that the fact about water is *entailed* (in Jackson's sense) by the microphysical facts (since it is metaphysically necessary that water is H_2O), but not that the entailment is a priori. To draw this further conclusion, we need a conceptual analysis that will mediate between claims about H_2O and claims about water. To illustrate the sort of conceptual analysis that might do this, Jackson invites us to "suppose that the right account of the semantics of 'water' is that it is a rigidified definite description meaning roughly 'stuff which actually falls from the sky, fills the oceans, is odourless and colourless, is essential for life, is called "water" by experts, . . ., or which satisfies enough of the foregoing'" (1993, 39). We will label the appropriate description, however the details are filled out, 'the waterish stuff', though in the context of the idea that the analyses are supposed to be functional (in terms of roles), we will use the terminology 'the water role'. Using this abbreviation and an "actual" operator that rigidifies a description, the definition of 'water' is as follows: water $=_{df}$ the actual waterish stuff.[3] Now, the statement 'H_2O = the waterish

[3]The rigidifying operator works as follows: the extension of 'the actual F', in any possible world w, is the thing that is the unique F in the actual world. In an earlier draft, we used some jargon from David Kaplan's work on demonstratives to represent rigidified descriptions: "dthat[the F]" instead of "the actual F." Kaplan's device is often interpreted as a rigidifying operator, but as Jim Pryor and Alex Byrne reminded us, this misrepresents Kaplan's original intentions in a way that is subtle, but important for the general issue we are discussing. As Kaplan first explained it, 'dthat' is not an operator on definite descriptions that turns them into rigid designators of the denotation of the description, but instead a semantically complete but context-dependent referring term. It is like a demonstrative 'that' whose use is accompanied by a pointing gesture, or a 'she' used in a context that somehow makes salient one particular female. The description in

stuff' is a *contingent* a posteriori truth. Let us assume for the moment that it can be deduced a priori from microphysics (with the help of other such definitions for the terms that occur in the definition—a complication we will also ignore for the moment). Given the definition, the statement 'water is the waterish stuff' is contingent, but a priori, since if we replace 'water' with its definition in this statement, the result will have the form *the actual F = the F.* (The reason 'the actual F = the F' is contingent is that it has a rigid designator on one side and a definite description on the other. In this it is like 'Grandma = the local bald thing'. Unlike the latter, however, the rigid designator is formed from the very definite description that appears on the other side, and that is what makes the identity a priori.)

From H_2O = the waterish stuff (microphysical truth) and water = the waterish stuff (a priori truth), it follows by logic alone that water = H_2O. So we can derive a priori from the two contingent premises

(a) the earth is covered 60 percent by H_2O
(b) H_2O = the waterish stuff

the contingent conclusion

(d) the earth is covered 60 percent by water.

The key is the a priori

(c) Water = the waterish stuff.

The strategy, in effect, is to factor the necessary a posteriori statement that water = H_2O into two parts, a contingent statement about H_2O that is (assumed to be) derivable from microphysics (H_2O = the odorless, etc. stuff) and a contingent a priori statement that can be justified by conceptual analysis (water = the odorless, etc. stuff).

Why can't the consciousness facts be shown to follow a priori from the microphysical facts in a parallel way? Both Jackson and Chalmers, like Levine, rely on conceivability arguments to cast

---

brackets following the 'dthat' should be understood, not as a constituent of the sentence, but as a substitute for the contextual features that do the job of fixing the reference—a bit of stage direction rather than a part of the dialogue. See Kaplan 1989, 578ff. for a discussion of the contrast between these two interpretations of 'dthat'.

doubt on the hypothesis that there could be analytic definitions, in physical terms, of the expressions we use to describe our phenomenal experience. There is, for example, Jackson's famous thought experiment about Mary, a neuroscientist who has been raised in a black and white room, and so has had no visual experience of colors such as red. She has learned all the relevant physical facts about color and about the physiology of color vision, but is still ignorant of what it is like to see red. It would be highly implausible to suppose that she might come to know what it is like to see red by engaging in some conceptual analysis and logical deduction. Then there are the philosophical zombies. Chalmers and others argue that there does not seem to be any contradiction in the concept of a zombie—a creature that is functionally and physically just like us, but that has no consciousness.

The considerations that Jackson and Chalmers appeal to can be used to argue directly for the metaphysical possibility of zombies, and so for the falsity of physicalism, bypassing any consideration of conceptual analysis or reductive explanation. But as Levine emphasized, the metaphysical intuitions are controversial. It is much less controversial to break the argument into two steps: first, to use the intuitions to support only the judgment that the concept of consciousness cannot be given an analytic definition in functional or physical terms, and second, to argue from this to the metaphysical conclusion, using the general thesis that conceptual independence implies metaphysical independence. We agree that, independently of this kind of argument, the hypothesis that zombies are metaphysically possible has considerable prima facie plausibility that needs to be explained away if the hypothesis is to be rejected. (See Nagel 1974 and Hill 1997 for attempts to explain it away that seem to us promising.) But our project here is not to explain away this controversial intuition, but to rebut the second stage of the argument that attempts to derive the hypothesis from the unanalyzability of the concept of consciousness.

4. A Reason for Doubting the Premise about A Priori Analysis

Jackson and Chalmers argue that a priori conceptual analysis in terms of basic physics (microphysics, in the terms that Chalmers uses and we will adopt) is required for a defense of the thesis of physicalism, and for the possibility of a physicalist explanation that

closes the explanatory gap. Since the concept of consciousness cannot be given such an analysis, the explanatory gap cannot be closed. But what is their *argument* that a priori conceptual analysis is required to close such a gap? What is offered is not an argument for this, but *examples* that show that *if* a conceptual analysis of a certain kind were always available, then we could use these conceptual analyses to account for the necessary a posteriori truths of reductive explanation. We have no quarrel with this conditional. What we doubt is that these conceptual analyses are very often available. They show that their model of reductive explanation *might* be right, not that it *must* be right.

Might a real explanatory gap be closed without a conceptual analysis of the terms in which the explanatory problem is posed? Consider how the explanatory gap was closed in the case of life. Famously, it was once thought that the explanation of life required appeal to some kind of vital force. How was this idea rejected? One might try to analyze life a priori in terms of reproduction, locomotion, digestion, excretion, respiration, and the like, and then give further analyses of these terms, eventually grounding the functions in microphysical terms. But one doesn't have to be a Quinean who rejects all a priori analyses to see that this one is hopeless. Note first that 'digestion', etc. are not terms of microphysics. In fact, not a single example of an analysis of a non-microphysical term in terms of microphysics is given by either Jackson or Chalmers. The most plausible candidates for a priori analyses are ones in which analysand and analysandum involve the same "family" of terms. Further, nothing in the concept of life rules out the possibility that there could be living beings that are immortal, and don't reproduce, that are tree-like (so don't locomote), get their energy by electromagnetic induction (so don't digest or excrete), and have no need for any substance in the air (so don't respire). Perhaps such possible beings take in chemicals from the soil but use all that they absorb so don't excrete any residue. Perhaps the definition will be of the reference-fixing kind: life $=_{df}$ the actual process that realizes the relevant cluster of functions. But is it really plausibly a priori that actual living things reproduce, locomote, digest, excrete, and respire (or any sufficiently large set of these)? For example, can't we imagine an alternative history in which physiologists discover that the humans don't do any digestion themselves—rather, our stomachs contain insects that digest our food

for us, excreting waste products which are exactly what we need to live.

More relevantly, it is doubtful that fulfilling any set of functions is conceptually *sufficient* for life. A moving van locomotes, processes fuel and oxygen, and excretes waste gasses. If one adds a miniaturized moving van factory in the rear, it reproduces. Add a TV camera, a computer, and a sophisticated self-guiding computer program, and the whole system could be made to have more sophistication, on many measures, than lots of living creatures.

These examples suggest that no a priori functional analysis has much to do with the closing of the explanatory gap in the case of life. Still, the explanation of how living creatures can carry out the functions used to characterize life is part of the story of how the gap was closed. For example, the discovery of the physical machinery of reproduction by Watson and Crick no doubt was a gap narrower. More generally, it seems reasonable to think that the story of the closing of the explanatory gap about life takes something like this form: There are some paradigm cases of living things, including some that are quite simple. (We need not assume that it is a *conceptual* truth that even the paradigm cases of living things are alive.) We understand completely how some of the simpler forms of life work. We have reason to think that more complicated living things work by similar principles, and see no bar, in principle, to our extending our explanations of simple living things to all forms of life. This particular model may not apply very well to consciousness, but the example of life does seem to us to illustrate how an explanatory gap can be closed without the sort of conceptual analysis that Jackson and Chalmers argue is required. Closing the explanatory gap in the case of life has nothing to do with any analytic definition of 'life', but rather is a matter of showing how living things around here work.

Jackson's and Chalmers's examples of conceptual analyses fall into two categories. Chalmers gives an example of reproduction as the production of a certain kind of copy. Perhaps a correct a priori functional analysis can be constructed along these lines. Chalmers does not fill in "certain kind," so we are given no reason to expect an analysis. But we very much doubt that any *analysis in microphysical terms* will be forthcoming.

The second category is that of a priori reference-fixing definitions, as with the example of 'water'. The functional analysis does

not give the meaning of the term in the ordinary sense, but is nevertheless supposed to be a priori and part of the "semantics" of the term and the concept, recoverable by anyone who understands the term, or possesses the concept. To rehearse Gareth Evans's example (1982, 31), to illustrate the pattern, we might use 'Julius' as an abbreviation for the actual inventor of the zipper. A conceptual analysis of the term 'Julius' that is available in principle to anyone who understands that term will reveal that Julius, if he exists, invented the zipper.

'Julius', of course, is a highly contrived example, as are all clear examples of such reference-fixing definitions. Even if descriptions play an essential role in fixing the reference of a name, there is no reason why they must become part of the semantics of the name in this way. The point has long been a familiar one; it was the main thesis of Kripke 1972 that ordinary proper names are *not* like 'Julius'. All of the examples of the second of Kripke's three lectures are directed at the refutation of a reference-fixing version of the description theory of names. Suppose we consider, not 'Julius', but 'Kevin'. This name was introduced—acquired its reference—in an act of dubbing that involved a definite description. The proud but pedantic new parent said, "Let 'Kevin' be the name of the baby in the third crib from the right." One might argue that Kevin's parent, at the moment of the act of dubbing, had a priori knowledge that Kevin was the baby in the third crib from the right (though we think even this is doubtful), but it is clear that the information is quickly lost, and is not something that can be extracted by competent users of the name from their understanding of it. Perhaps the semantics of 'water' is more like 'Kevin' than it is like 'Julius', in which case there is no way to fill in the details of 'the water role' so that it is a conceptual truth that water occupies the water role. And of course it is even more doubtful that any such analysis of the water role would be both a conceptual truth and be an analysis in *microphysical* terms.

In sum, we have strong reason to doubt that reductive explanations that close explanatory gaps depend on the kind of a priori conceptual analyses that Levine, Jackson, and Chalmers appeal to.

5. Uniqueness

Let us look a bit more closely at the supposed conceptual truth that water = the waterish stuff. Recall that 'the waterish stuff' ab-

breviates an appropriate cluster of descriptions of the superficial properties of water by which it is commonly identified by competent users of the term, something like "the stuff that falls from the sky, fills the oceans, is odourless and colourless, is essential for life, is called 'water' by experts . . ., or which satisfies enough of the foregoing." Now the assumption that H_2O is *a* waterish stuff is not enough to ensure that H_2O = water, for there might be other waterish stuffs too. For Jackson's pattern of argument to go through, H_2O must be *the* unique waterish stuff. But we don't want our definition to rule out the possibility that there are other waterish stuffs elsewhere that are unrelated to our applications of the concept of water, for example, XYZ on twin earth (where here, twin earth is not regarded as a counterfactual possibility, but as a distant part of our universe). So the water role must include an indexical element, which we can specify by adding 'around here'. Then our definition of 'water' is: water = the actual (unique) waterish stuff around here. (The indexical condition is independent of the rigidifying operator.) And the relevant conceptual truth will be a version of (c):

(c) water = the (unique) waterish stuff around here.

Then the overall argument says that the two microphysical facts, (a) that H_2O covers 60 percent of the globe, and (b) that H_2O = the waterish stuff around here, together with the conceptual truth (c), logically imply (d) that water covers 60 percent of the globe. Since (c) is a conceptual truth, the entailment of (d) by (a) and (b) can be established a priori. That is, since it is a conceptual truth that water = the actual waterish stuff around here, it is also a conceptual truth that water = the waterish stuff around here. So from (b) we can move a priori to: water = H_2O, and from (a) and (b), we can move a priori to (d).

We doubt that there is a way to spell out 'waterish' so that (c) is a conceptual truth, and we doubt further that any remotely plausible analysis would get to first base when substituted in (b) as a truth of microphysics; but even if we set this issue aside, there is an additional problem with the argument: Even if 'water' can be given such a definition, the uniqueness assumption in the description throws doubt on the claim that premise (b) of the argument is a microphysical fact.

If 'waterish' is spelled out so that it is a conceptual truth that

only a physical stuff can be waterish, then it will perhaps be a microphysical fact that there is only one waterish stuff around here. But for at least some names for substances or properties that are in fact physical, the reference-fixing definition might be a functional one that did not exclude on conceptual grounds the possibility that the substance or property be nonphysical. Consider a different example: at one time, heat was thought to be a fluid substance, caloric. Imagine an alternative scientific history in which the caloric theory coexisted with the observation that molecular kinetic energy goes up and down, covarying with influx and outflow of heat. The theory might have arisen that what we might ahistorically call "ghost heat" was correlated with but not identical to molecular kinetic energy. Let us define 'heatish' by analogy to 'waterish'. A heatish stuff satisfies most of a list of properties such as: produces sensations of warmth,[4] makes water boil when added, makes water freeze when taken away, is called 'heat' by experts. Then the comparable claim to the one being discussed would be: the (unique) heatish stuff around here is mean molecular kinetic energy. But the claim that mean molecular kinetic energy = the (unique) heatish stuff around here is not a purely microphysical claim, since it rules out the possibility that ghost heat is also a heatish stuff around here.

Applying this point to Jackson's argument, we can't move a priori from microphysics to the claim that water covers 60 percent of the earth because there is a world (considered as actual) in which some of what counts as water is ghost water, so even if H_2O covers 60 percent of the earth in that world, what they call 'water' will cover more. Further, we can't move a priori to 'H_2O covers *at least* 60 percent of the earth' because ghost substances could increase the surface area of the earth, or alternatively, cover some of the H_2O, putting it below the surface.[5]

[4]We will ignore the fact that reference-fixing definitions of terms like 'heat' might be expected to contain such a reference to sensations.

[5]Chalmers acknowledges a "minor complication" (1996, 40), that "negative facts" do not follow a priori from microphysics. A positive fact in W is defined as a fact that holds in every world that contains W as a proper part. We are skeptical about the use of 'proper part' as a relation between worlds. We understand possible worlds as ways things might be. What is a proper part of a way things might be? To the extent that we get a grip on the notion, we doubt that there are any positive facts.

Recall that Jackson's definition of physicalism recognized the possibility of a world that is a microphysical duplicate of the actual world, while also containing some additional nonphysical substances and properties. In response, a "nothing but" condition is built into the definition of physicalism by requiring for the truth of physicalism only that *minimal* physical duplicates of the actual world be duplicates *simpliciter*. But the need for this qualification presupposes that the "nothing but" condition is not something that is entailed by microphysics *itself*; that is, *it is not a claim of microphysics that our world is a minimal physical duplicate of itself*. To take the "nothing but" condition to be an implicit claim of microphysics would be *to build the thesis of physicalism into microphysics*, which philosophers such as Jackson and Chalmers, who reject physicalism, should be reluctant to do. They reject physicalism, not microphysics. The truth of what physicists write in textbooks does not depend on the mind-body problem. Even if it is a microphysical fact that H_2O is *a* waterish stuff around here, it is not a microphysical fact that it is *the* waterish stuff around here.

Reacting to this point, Jackson remarked (personal communication) that the medical profession would be outraged at the idea that they have failed to show that fairies don't cause cancer. True enough. We also think that the physics profession has sufficient reason to rule out the existence of ghost heat. But we are making a point, not about what it is reasonable for scientists to believe, but about what can be deduced a priori from their theories. The kind of causal overdetermination that one would have to hypothesize to reconcile either ghost heat or fairies that could cause cancer with accepted theories would have not a shred of scientific plausibility, but this judgment is quite compatible with the claim that there is no logical derivation from medical science (or thermodynamics), together with conceptual truths, of the absence of such nonphysical overdetermining causes.

Our hypotheses about "ghost" properties—redundant nonphysical phenomena—are obviously contrived, but they are relevant to the issue, since on the dualist view of consciousness that Jackson and Chalmers support, consciousness is a kind of ghost property—something that is either epiphenomenal or a property with redundant overdetermining influence on the physical world. We argue that hypotheses about ghost properties are ruled out on empirical methodological grounds, rather than by conceptual analysis. The

same kind of methodological considerations might be used to argue against dualism about consciousness.

In section 1, we noted the distinction between what can be deduced a priori from microphysics alone and what can be deduced a priori from microphysics augmented by the claim that certain microphysical entities are the referents of *our* words. We noted that one cannot move a priori from microphysics alone to the conclusion that H_2O boils, since one might not know that water is H_2O. If one's 'water' picked out XYZ instead of H_2O, then one's 'boil' would pick out a kind of thing that happens to XYZ when it is heated enough rather than the superficially similar behavior of sufficiently heated H_2O. This point also applies in the current context. One cannot move a priori from microphysics alone to the claim that there is water around here. The problem cannot be evaded by appeal to the formula "Add indexicals to microphysics." There might be both H_2O *and* XYZ around here. What has to be added to microphysics is something that has the upshot that our term 'water' refers to H_2O, and it is not clear that this can be done without simply assuming that the microphysical facts determine the referential facts, an assumption that is no part of microphysics.

We have argued that the assumption that there is at most one waterish stuff around here is essential to the argument, and that it cannot be extracted from microphysics. The argument might still work if we could deduce this proposition instead from the alleged conceptual truth (c) water = *the* waterish stuff around here. As noted above, when 'water' is replaced by its definition, (c) is seen to have the form 'the actual F = the F', a form of a statement that is said to be knowable a priori, but nevertheless to express a contingent proposition. Consider the analogy mentioned earlier to 'Grandma = the local bald thing'. This is obviously contingent, since Grandma might not have been bald and someone else might have been the local bald thing. But 'the actual local bald thing', like 'Grandma', is a rigid designator naming Grandma. So it is just as contingent that the actual local bald thing = the local bald thing. Though contingent, it is also a priori, since the rigid designator is formed from the same definite description that appears on both sides of the identity sign. If it designates anything, both occurrences designate the same thing. But what if, in fact, there is no unique local bald thing? Then the reference fixing fails, and no contingent proposition is expressed. In this case, what is the

status of the statement? We might say that when uniqueness in fact fails, the statement that the actual F = the F is not true, since it expresses no proposition at all. All that is really a priori is this: *if* there is a unique F, then the actual F = the F. Alternatively, we might stipulate that 'the actual F = the F' (and therefore (c)) shall be true anyway in this case. But then, although the statement can be known a priori, it doesn't imply that there is a unique F. The upshot is that if, in Jackson's derivation, we replace (b), which we have argued is not a microphysical claim, with the weaker claim that H_2O is *a* waterish stuff around here, then there will be no way to derive the conclusion from the premises. Even if it is a priori that water, *if there is such a thing*, is the unique waterish stuff around here, that won't suffice to get out of microphysics the conclusion that water covers 60 percent of the earth.

The possibility of additional waterish stuff that is nonphysical is not, perhaps, to be taken too seriously, but it is not so implausible to imagine the discovery that even though water was believed to be the unique waterish stuff, in fact there turn out to be two or more different (physical) kinds of waterish stuff. What should we say in such a case? If it is a definitional truth that water = the actual unique waterish stuff around here, then we should have to conclude that there is no such thing as water, and set about to construct a new definition. This is one possible response, but there are others, and choosing between them seems to be a decision that will be driven by empirical and theoretical considerations. We might decide that the kind term partially denotes both occupants (see Field 1973 on 'mass'). Another possibility is that we should regard the term as a non-kind term denoting a disjunctive property, the disjunction of the two items. A third possibility is that we should regard the kind term as denoting the *role* property shared by the two items rather than the *fillers* of the role. In any actual case, the matter may be decided by the relevant details. It is a part of the semantics of natural kind terms that they are natural kinds, but it may also be part of the semantics of these terms that this is a defeasible condition. What is not plausibly part of the semantics, something we all know in virtue of knowing our language alone, is what to say in all the myriad cases in which the defeasible condition is defeated. In these cases, what we should say will no doubt be dictated by principles of "simplicity," conservativeness, etc.

These issues often arise in medicine where disease names are

used to denote the filler of a role on the assumption that there is a single filler. Then it is discovered that there are many fillers. For example, the term 'rheumatism' was once used to mean a disease characterized by pain in the joints or muscles that was caused by the flow of an evil fluid which flowed from the brain to the affected parts of the body. The term comes from the Greek 'rheuma', meaning a watery discharge, but the use of the term described here dates from the seventeenth century. The term was used for very different diseases, including rheumatic fever, arthrosis, and arthritis (which is itself a term that covers different diseases, some caused by infection, some by autoimmune problems, and another by wear and tear). As it slowly dawned on the medical community that the cases lumped under that term had nothing in common but the symptoms, the name was used to denote the syndrome itself. Many disease names—'lumbago', for example—follow this course.

The contrast mentioned above between a disjunctive kind and a role property is important to the point we are making about uniqueness. If we took 'jade' to denote a disjunctive non-kind, we would take it to denote the disjunction of jadeite and nephrite, the two types of jade. If instead we took 'jade' to denote a role property, we would take it to denote a cluster of superficial properties such as a certain color, weight, hardness, shapeability, and the like. Or perhaps it should be construed as the property of having most of the elements of the cluster. What is it for something to be jade? Is it to have the property of being either jadeite or nephrite? Or is it to have a certain color (greenish whitish), etc.? We are not sure. If we discovered a new substance that had those superficial properties, wouldn't that be a new kind of jade? If so, or if it is indeterminate, then the disjunctive analysis is not right.

Our point doesn't require making up our minds about jade or disease names. All we need is these two assumptions: first, there is a real distinction between disjunctive properties, such as the property of being either jadeite or nephrite, and superficial cluster properties, such as the property of having a certain weight, hardness, color, etc.; second, in some cases in which the paradigm examples of some putative natural kind term turn out to be members of two or more quite different natural kinds, it would be right to take the term to denote the superficial property, or to have indeterminate reference. If it could, for all that is built into the con-

cept, have turned out that 'water' referred to any waterish stuff, or that 'water' had indeterminate reference, then this is enough to show that we should not regard it as a piece of conceptual analysis that water is the actual waterish stuff around here.

We have been criticizing the idea that there is any way to specify an appropriate meaning for 'waterish' or 'heatish' so that it is a conceptual truth that water = the actual unique waterish stuff around here or that heat = the actual unique heatish stuff around here. The criticism has depended on the issue of what we should think if there turns out to be two different role fillers. But the fact that these are not conceptual truths can be seen another way, by a consideration of *why* we usually think that there is only one filler. The supposition that it is a conceptual truth that heat = the actual unique heatish stuff around here is incompatible with the actual practice of scientific reduction. The claim that heat and molecular kinetic energy are dual occupants of the same role is not false because it falls afoul of the concept of heat. The view that heat and molecular kinetic energy are two rather than one is not contradictory or conceptually incoherent. It is false, and can be shown to be false by attention to certain methodological principles whose power and importance are widely acknowledged even if no one has ever been able to formulate them precisely. We refer to the set of methodological principles that are usually invoked with the misleading name 'simplicity'.

6. Digression on "Simplicity"

Levine, Jackson, and Chalmers suppose that the gap between descriptions in terms of microphysics and descriptions in terms of, for example, 'water' and 'heat' is filled by conceptual analysis. A deep inadequacy in this view is revealed by the role of methodological considerations in our actual decisions about such matters. Why do we suppose that heat = molecular kinetic energy? Consider the explanation given above of why heating water makes it boil. Suppose that heat = molecular kinetic energy, pressure = molecular momentum transfer, and boiling = a certain kind of molecular motion. (We are alluding to an empirical identity claim, not the a priori behavioral analysis considered earlier.) Then we have an account of how heating water produces boiling. If we were to accept mere correlations instead of identities, we would only have an ac-

count of how something correlated with heating causes something correlated with boiling. Further, we may wish to know how it is that increasing the molecular kinetic energy of a packet of water causes boiling. Identities allow a transfer of explanatory and causal force not allowed by mere correlations. Assuming that heat = mke, that pressure = molecular momentum transfer, etc. allows us to explain facts that we could not otherwise explain. Thus, we are justified by the principle of inference to the best explanation in inferring that these identities are true.

If we believe that heat is correlated with but not identical to molecular kinetic energy, we should regard as legitimate the question of why the correlation exists and what its mechanism is. But once we realize that heat *is* molecular kinetic energy, questions like this will be seen as wrongheaded.

Suppose one group of historians of the distant future studies Mark Twain and another studies Samuel Clemens. They happen to sit at the same table at a meeting of the American Historical Association. A briefcase falls open, a list of the events in the life of Mark Twain tumbles out and is picked up by a student of the life of Samuel Clemens. "My Lord," he says, "the events in the life of Mark Twain are exactly the same as the events in the life of Samuel Clemens. What could explain this amazing coincidence?" The answer, someone observes, is that Mark Twain = Samuel Clemens.[6] Note that it makes sense to ask for an explanation of the correlation between the two sets of events. But it does not make the same kind of sense to ask for an explanation of the identity. Identities don't *have* explanations (though of course there are explanations of how the two terms can denote the same thing). The role of identities is to disallow some questions and allow others.

This point about identities is relevant not only to identity theorists, such as Smart, who identify mental properties with physiological properties of the brain, but also to functionalists who identify mental properties with functional properties, arguing that they are realized by, rather than identical with, physiological or (ultimately) microphysical properties. The functionalist makes identity claims too—that mental properties are functional properties—and we ar-

[6]One of us heard this story somewhere, but we don't know where (see Block 1978).

gue that such claims might be a posteriori claims that are justified on methodological rather than conceptual grounds.

Smart (1959) asked the question of what shows that pain is identical to a brain process as opposed to merely correlated with it. He invoked methodological considerations of the sort mentioned above (avoiding "nomological danglers," he said). Let's lump such methodological considerations together under the unfortunate heading 'simplicity'. In a paper that inspired Jackson and Chalmers, Lewis (1966) argued that simplicity was not needed—our concepts are enough. The considerations mentioned here favor Smart over Lewis.

7. A Different Kind of Analysis?

We have been arguing against the plausibility of a certain kind of conceptual analysis of terms like 'water' and 'heat'. Jackson and Chalmers would reply, we suspect, that they have no commitment to any particular conceptual analysis, and so their general point is not affected by arguments against the plausibility of any particular analysis, or pattern of analysis. But, they would insist, there *must* be some kind of analysis. "Of course any view about how 'water' gets to pick out what it does will be controversial. But it is incredible that there is *no* story to tell. It is not a bit of magic that 'water' picks out what it does" (Jackson 1993, 42 n. 25). Our reply is that of course there must be a story to be told (whether physicalism is true or not) about how 'water' comes to refer to water, but we need not assume that the story involves a conceptual analysis in microphysical terms, or that the story is one that is available a priori to all competent users of the word 'water'. We suggested above that the story to be told might be the kind of account that Kripke argued should be given for proper names; Jackson suggests, at one point, that his argument could accommodate such a story—that it is just another kind of conceptual analysis. "If you prefer a causal-historical theory," he says, "you would have to replace [the second premise, (b), of the argument] by something like 'It is H_2O that was (the right kind of) causal origin of our use of the word "water"'" (1993, 42 n. 25).

According to this suggestion, the argument pattern would now be something like this:

(a) 60 percent of the earth is covered with H_2O.

(b′) H_2O is the stuff that plays the right kind of causal role in explaining our use of the word 'water'.
(d) Therefore, 60 percent of the earth is covered with water.

The assumption must be that the inference from (a) and (b′) to (d) is now mediated by the following conceptual analysis:

(c′) Water is the stuff that plays the right kind of causal role in explaining our use of the word 'water'.

Even granting, for the moment, that (b′) is a truth of microphysics, there is a problem: (c′), of course, is not a conceptual analysis of 'water'. It conveys no information about the meaning of the word 'water' that distinguishes that word from any other that names a liquid. If there is a gesture toward a conceptual analysis of anything in the allegedly a priori (c′), it is an analysis of reference, or meaning, and it would be this only if the phrase "the right kind of causal role" is taken (as we assume Jackson intends) as shorthand to be filled in with a substantive account of what the right kind of causal role is. But what reason is there to think that the story to be told about the relation that constitutes reference or meaning is a conceptual analysis (in microphysical terms), or is true a priori?[7]

Suppose we were to grant that there must be an *a priori* conceptual analysis of reference in physical terms to be found. Then consider the following argument pattern:

[A] Pyramidal cell activity is taking place in Jones at time t.
[B] Pyramidal cell activity is the process that plays the right kind of causal role in explaining our use of the word 'pain'.
[D] Therefore, pain is taking place in Jones at time t.

The inference from [A] and [B] to [D] is justified by the following quite trivial instantiation of our account of reference:

[C] Pain is the process that plays the right kind of causal role in our use of the word 'pain'.

Of course, the defender of physicalism must defend the crucial

[7]Cf. Kripke 1972, 88 n. 38, where Kripke comments on a point made by Robert Nozick that, in a sense, a description theory of names would be trivially true if there were a reductive analysis of reference of any kind. But Kripke disclaims any attempt to give such a reductive analysis, and expresses doubt that there is one to be had.

premise [B], but this is a purely contingent, entirely physicalist claim—at least, it is if (b′) is. The fact, if it is a fact, that we can imagine pain without pyramidal cell activity or pyramidal cell activity without pain, seems quite irrelevant to the defense of [B], and since the conceptual analysis is an analysis of reference, it is not clear that these possibilities are relevant to the truth of [C] either. It may be said that epiphenomenalism shows that [C] is not an a priori truth, but has no such impact on (c′). But if epiphenomenalism is coherent, we don't see why the comparable doctrine with respect to water or heat is not coherent. Suppose, for example, that we refer to pains via a "causal fork"—a brain state causes both the pain and our uses of the word 'pain'. Why should an analogous claim for 'water' be ruled out a priori? [C] and (c′) stand or fall together.

So while we agree with Jackson that there must be some nonmagical story to be told about how our words manage to refer, we don't agree that such a story will drive a wedge between reductionism about water and reductionism about pain.

8. Can Jackson and Chalmers Make Do with Approximate Conceptual Analyses?

Jackson emphasizes that there is an element of stipulation involved in the a priori entailments from physics of everything outside of consciousness. In discussing the example of the entailment of the facts of solidity from physics, he says, "That's what it takes, according to our concept, to be solid. Or at least it is near enough" (1993, 103). Near enough for what? His answer is: near enough for practical purposes. "For our day to day traffic with objects, it is the mutual exclusion that matters [as opposed to being everywhere dense], and accordingly it is entirely reasonable to rule that mutual exclusion is enough for solidity" (107).

But it can hardly be assumed that what matters for practical purposes is a priori. What matters for practical purposes depends on the facts of psychology and economics (for example). That is why the agents in Rawls's "original position" need to know basic facts of psychology and economics. Introducing such a notion of approximation does not seem a promising way to avoid the problems we have been raising for the a priori entailment from physics of everything outside consciousness.

9. The Upshot

To sum up, we have argued:

> There is no reason to believe that reductive explanation of facts about phenomena described in our ordinary folk vocabulary requires conceptual analyses of the terms of that vocabulary in microphysical terms, and there is no reason to believe that there are often such analyses to be found.
>
> Even if one had such an analysis (of the form "water is the unique waterish stuff"), another assumption of the argument would still fail, since the claim that H_2O is *the* unique waterish stuff would not be a microphysical claim. Given the possibility of ghost water that covers part of the earth not covered by physical water, it cannot follow from microphysics that water covers 60 percent of the earth. (Note that the ghost water point is not directed at the claim that there are a priori conceptual analyses of ordinary folk terms in microphysical terms. Chalmers and Jackson say little to support that claim and we have given some reason to doubt it. We use the ghost examples to show that even accepting the conceptual analyses, the water facts do not follow a priori.)
>
> It cannot be a priori that water = the occupant of the such and such role, since in some circumstances the right thing to say is that a term picks out the role itself—or is indeterminate between the role and its occupant. And it remains to be shown that there is some more complicated analysis (for example, a set of conditionals) that is really a priori.
>
> The claim that water is *a* waterish stuff might be a microphysical claim, but if Jackson and Chalmers try to make do with this weaker premise, their argument faces the following dilemma: the allegedly a priori conceptual truth that water is the unique waterish stuff either has an existential presupposition (that there is a unique waterish stuff), in which case it is sufficient for the argument, but not knowable a priori, or else it does not imply the existential claim, in which case it may be a priori, but will be insufficient for the argument.

Suppose it were agreed that no conceptual analyses for natural kind terms like 'water' and 'heat' are available, and so that there

is no a priori entailment of statements about water by microphysical theory and fact. What would the significance of this be? Why would it matter for dualism and the explanatory gap? Does it really challenge the overall drift of the Levine-Jackson-Chalmers point of view?

The point of view that we are criticizing depends on two claims that we are accepting: first, that Jackson's Mary cannot deduce the facts about what it is like to see red from the microphysical and functional facts; second, that there is no contradiction or incoherence in the extreme zombie hypothesis, the idea of a microphysical duplicate of one of us but with no consciousness—that is, even with a full knowledge of microphysics and the microphysical state of a conscious human being, one could not come to know, by reasoning from these facts, that the person was conscious. These ideas are supposed to show that the facts of consciousness are not *a priori* entailed by the microphysical facts. But we have been arguing that the facts about *water are not a priori entailed by the microphysical facts either.* To derive 'The earth is 60 percent covered by water' from microphysics, we need the a posteriori (necessary) truth that water = H_2O. But now we can see that even if Jackson's and Chalmers's thought experiments show something about what can be extracted a priori from microphysics, they don't show any disanalogy between the concept of consciousness and the concept of water. In particular, they don't show that an identity claim connecting consciousness with a kind of neurophysiological process cannot be true, and cannot play a role in closing an explanatory gap. (Note that we are not saying that the identity closes the gap all by itself.) Suppose that there is an a posteriori truth of the form consciousness = brain state B. Recall the suggestion by Crick and Koch, mentioned earlier, that consciousness is pyramidal cell activity. Then the facts of consciousness are necessitated by the microphysical facts (though not entailed a priori). The fact that the notion of an unconscious zombie is not a contradiction cuts no ice.

Without the help of a conceptual analysis, how might such an identity claim be justified? By using the kinds of methodological consideration sketched in our discussion of simplicity above—the same kinds of considerations that are used to justify the claim that water = H_2O. We said that simplicity considerations are not a priori, but this is not really necessary to the argument. Let simplicity be as a priori as you like. Whatever a priori status simplicity

has will apply as well to reductive explanation of consciousness as to reductive explanation of water.

After making the microphysical world, did God have to add consciousness? Not if the identity claim is true, and we might have reason to believe that it is, even without the help of a conceptual analysis.

10. The Two-Dimensional Framework

The conceptual analysis thesis, in the form that we have been discussing it, requires explicit a priori verbal analyses in microphysical terms (or at least a priori sufficient conditions in microphysical terms) of all concepts that do not involve consciousness. This is an incredible claim. The best candidates for a priori conceptual analyses have been ones in which the terms of the analysis are in the same "family" as the term analyzed—for example, one mentalistic, epistemic, and normative concept is analyzed in terms of others. But the conceptual analyses we are criticizing would analyze ordinary terms in terms of a language (that of microphysics) that is unknown to ordinary people.

However, there is no real commitment on the part of the philosophers whose views we are criticizing to any such analyses. The official view stated by Jackson and Chalmers is not that the relevant concepts can really be given verbal analyses but that they have meanings or intensions of a particular kind. To spell out their account, they appeal to the well-known two-dimensional framework. We will argue that this appeal misfires. The two-dimensional account does nothing at all to motivate the claim that there is an a priori accessible conceptual component of content. If there is such a component, it can be represented in terms of the two-dimensional apparatus, but that apparatus provides not the slightest reason to believe in it. Further, even if there is such a component, it cannot be used to drive a wedge between consciousness and realms where physicalism is not controversial.

To motivate the two-dimensional apparatus, we will begin by considering some familiar externalist objections to the kind of specification of *the water role* that we have been considering. It is argued that we are not guaranteed a priori that water is the—or even *an*—odorless drinkable liquid in rivers and lakes that we have been calling 'water', since it is not even guaranteed a priori that water

is a liquid. There is a twin earth in which the stuff that they call 'water' is H_2O, as here, but the stuff that they call 'liquid' is virtually all a slippery granular solid (White 1982). According to this story, water is an exception, one that the residents of Twin Earth would not call 'a liquid' if they knew the scientific facts. (Imagine that on Twin Earth, water is rare.) Because (the argument goes) the counterfactual situation in which this story is true is a twin of ours, an utterance is actually a priori only if the counterpart utterance in that situation is a priori there. Since the counterpart of "water is a liquid" is false there, it cannot be a priori true.

Note that the uses of 'liquid', 'solid', etc. that are assumed here are "natural kind" uses. There are nonnatural kind uses which for lack of a better term we will call "superficial" uses of these terms according to which on the twin earth specified above, water is correctly called 'a liquid'. This fact may be used to motivate a reply, namely that the sense of 'liquid' assumed in the reference-fixing definition is the superficial sense. What is this superficial sense? The idea is that just as 'water' is associated with a cluster of superficial properties that some other stuff could possibly have, so 'liquid' is associated with superficial manifestations (which we might abbreviate with 'liquidish', or 'the liquid role'). A spelling out of the liquid role can be expected to reveal further terms that have an externalist dimension, since as Tyler Burge and others have persuasively argued, the dependence of meaning and content on an external environment is not restricted just to natural kind terms, but is pervasive, extending even to color terms (Burge 1978). The issue about how deep externalism goes is controversial, and familiar from the extensive literature on narrow content. We will not try to settle it here, since our point is just to motivate a strategy for bypassing the question by trying to give a general method for abstracting away from the external component of content without giving any particular conceptual analyses, yielding a general representation of the information that can be extracted from concepts by understanding alone.

Grant that the meaning of a word, or of a concept,[8] has an ex-

[8]It is a familiar point that the word 'concept' is ambiguous, and we think that equivocation on it is implicated in some of the confusions in the application of the two-dimensional apparatus. Sometimes 'concept' refers to something abstract like a meaning. Concepts are what predicates are used to express, as propositions are what sentences are used to say. Propositions are also the contents of thoughts (acts of thinking, states of be-

ternal dimension; that is, grant that the fact that a word or concept *has* a certain meaning is not a fact wholly about the internal state of the speaker or thinker, and is not a fact that is available a priori to everyone who understands the word or possesses the concept. This implies that the extension of one's concepts, and the truth values of one's thoughts, may depend on external facts in two different ways. First, external facts contribute to determining what their content is, and second, they contribute to determining whether and where a concept with that content is instantiated. That is, first there are the facts, whatever they are, that make it true that 'water' refers to water (H_2O), and second, there are facts such as that there is water (H_2O) covering the basement floor. Together, these facts imply that the thought that one would express with the sentence 'There is water all over the basement floor' is true. Now even if the externalist disputes the possibility of conceptual analysis, he cannot dispute that extensions and truth values depend on the facts in these two different ways. The strategy of the two-dimensional analysis is to use this distinction to separate out the purely conceptual component of content.

The world is all that is the case, and a possible world is all that would be the case if that world were actual. So whatever external (or internal) facts are relevant to the determination of the content of an expression or a concept, a specification of a possible world will include a specification of those facts. So we can identify the purely internal, purely conceptual component of content with a function from possible worlds (or possible worlds centered on a speaker or thinker and a time) to contents. The value of the function—the content in the ordinary sense of the relevant concept or expression—may itself be identified with a function from possible worlds to extensions, so the conceptual component of content will be a *two-dimensional intension*—a function from (centered) possible

---

lieving); concepts, in this semantic sense, are perhaps components of what is said and thought. Other times, by 'concept' one is attempting to refer to a mental analogue of a predicate—a mental word rather than what the mental word expresses. If one applies the type-token distinction to concepts, or if one talks about the content of a concept, one is using the word in this sense. It is controversial whether there are concepts in this latter sense and whether if there are, they play any significant role in thought, but we will go along for the ride, following Chalmers in using 'concept' in this sense.

worlds to functions from possible worlds to extensions (truth values, in the case of sentences). Or equivalently, it will be a function taking two arguments—a centered world and a world—into an extension.

The reason that the first argument of the function is a centered world is that one and the same utterance type can occur in different contexts within the same possible world. For example, water and twin water can exist in the same world; 'water' in one mouth can pick out water and in another mouth pick out twin water. So the two-dimensional intensions must be functions whose arguments are more fine-grained than possible worlds: a centered world is a world plus a designated spatiotemporal location in the world.[9]

Given the two-dimensional intension, we can define two other kinds of intensions, which in Chalmers's terminology are called the primary and secondary intensions. The secondary intension is the ordinary intension, the one to which the externalist arguments apply. It is the function from worlds to extensions that are the values of the two-dimensional function. The primary intension, like the secondary intension, takes just one argument into extensions, though in this case the argument is a centered world. The value of the primary intension for a given centered world as argument is defined to be the same as the value of the two-dimensional intension for the pair of arguments consisting of the centered world and the same world, uncentered. Slightly more intuitively, to get the value of the primary intension in a given world, ask what the extension would be in that world considered as actual—what the words *as used in that world* refer to in that world. If on counterfactual twin earth, there is XYZ (but no H_2O) in the relevant bathtub, then we say (speaking in the actual world) that it is false that in that counterfactual world there is water in the tub. But our twin there truly says, "There is water in the tub." And if we associate the same two-dimensional intension with their utterance as with ours, then using Chalmers's terminology, we say that the secondary intension (determined in the actual world) is a proposition that is

[9]In other applications of this kind of framework, this complication is not necessary since if it is a particular token utterance, rather than an utterance type, that is associated with a two-dimensional intension, then both arguments of the function can be simply possible worlds.

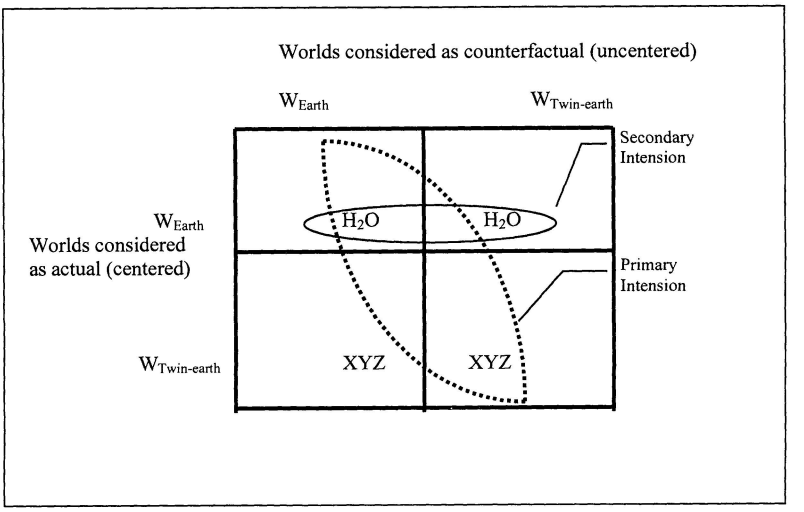

Fig. 1. 2-D intension of 'water'

false on counterfactual twin earth, while the primary intension is a proposition that is true there.

A speaker or thinker who knows what the two-dimensional intension of some expression is, and so who knows what the primary intension is, might still be ignorant of the secondary intension, which unlike the primary intension may vary from world to world. It is this variation that represents the dependence on external facts to which the externalist points. But the primary intension is determined (by the two-dimensional intension) independently of any particular possible world, and is supposed to represent the internal component of content, the component that can be determined a priori.[10]

[10]The abstract two-dimensional semantic apparatus has its origin in work by Hans Kamp and Frank Vlach, extended and applied by David Kaplan in his theory of demonstratives. An important purely abstract development of two-dimensional semantics is Segerberg 1973. Stalnaker 1978 is a more general application of the framework to semantic and pragmatic phenomena, including the use of it to represent Kripke's distinctions between a priori/a posteriori and necessary/contingent. Davies and Humberstone 1980, another application of the apparatus, is the source of the terminology "world considered as actual" versus "world considered as counterfactual." White 1982 applies a version of Kaplan's content/character distinction to the language of thought. Stalnaker 1990 and Block

So without worrying about how to specify its details, we can represent the water role simply by identifying it with the primary intension of 'water'. All the problems that we have raised for conceptual analysis may now seem to disappear. First, recall our example of a counterfactual twin earth in which our duplicates use 'water' to pick out water, but use 'liquid' to pick out a type of substance (a slippery granular solid) that doesn't include water. This is a counterexample to any definition of water that includes the requirement that it be a liquid, and to the assumption that it is knowable a priori that water is a liquid. But the fact that water is not correctly called 'a liquid' in this possible world does not prevent the primary intension of 'water' from picking out water there. All it shows is that the primary intension of 'liquid' may fail to apply to something to which the primary intension of 'water' applies. A similar point can be made in response to the type of consideration advanced by Burge (1978). Imagine a twin earth in which you exist just as you are internally but in which the words 'colorless', 'odorless', 'river', 'lake', and 'thirst-quenching' are all used differently by your language community so as to make all of these words fail to apply to water. Any of the familiar reference-fixing definitions would be false (showing that they are not really a priori), but that needn't keep your word 'water' from picking out water. If it does, then the (actual) primary intension of 'water' still picks out water in that counterfactual world. The coherence of such twin earth scenarios might be disputed, but with the two-dimensional apparatus, one can bypass that issue. The proposal that we identify the contents of concepts with primary intensions avoids the problem of the rigidified description theory by not *defining* 'water' in words. The primary intension is just a function from centered worlds to extensions. So the primary intension for our term 'water' maps each world into the right stuff, whether or not it is a liquid, and whether or not the linguistic community uses words differently from the way the utterer of 'water = H_2O' uses them. The right stuff is the stuff that is properly related to our uses and those of our language community. In any world in which our word 'water' actually picks out something, the primary intension will yield that very value (cf. White 1982).

1991 consider the use of the apparatus to define a notion of narrow content.

Further, the primary intension apparatus appears to get around the problems of uniqueness that we discussed above. If in fact there is one thing that satisfies the primary intension for (our actual word) 'water' in the actual world, it is water. That is an a priori conceptual truth, for after all the primary intension is simply defined to ensure that this is true. Of course, some worlds have no water at all, and some worlds have a number of different substances that have an equal claim to the water role, as with jade in the actual world. As we suggested above, in some of the worlds in which it is discovered that there are several kinds of stuff to which the term 'water' has been applied, it would be decided by rational and knowledgeable speakers that there was no such thing as water; in other such worlds, they would decide that water was a disjunctive kind; and in still others, they would decide that water, despite what was previously thought, was really a superficial kind, the term applying to whatever has certain superficial properties commonly associated with water. Since the primary intension (of *our* word 'water' in the actual world) is determined by what rational and appropriately informed speakers would say in the different possible worlds, this intension will therefore, by definition, account for any of these possibilities.

So, it seems, none of the objections we have raised to reference-fixing conceptual analyses apply to primary intensions. So let us just identify the water role with the primary intension of 'water', and use that as our definition in Jackson's pattern of argument. Reflection on the primary intension together with the microphysical facts about this world is enough to determine that H_2O satisfies that primary intension, that is, that H_2O = the unique waterish stuff. So we have achieved the Holy Grail: it is an a priori conceptual truth that if H_2O covers 60 percent of the globe and H_2O satisfies the 'water'-primary intension, then water covers 60 percent of the globe!

We hope that the reader is by now a little suspicious. How can so little do so much? Have we really succeeded, by this simple and perfectly general abstract maneuver, in identifying and isolating, for any expression, a component of its content that both is accessible a priori to anyone who understands the expression and will do the required work in Jackson's argument? We will argue that this apparatus contributes nothing to the identification of a purely conceptual or a priori knowable component of the content of a

concept, or to the support of a claim that there is such a thing to be identified. At best, it provides a *framework for representing such a component*, should there be one to be represented. And we will also argue that even if it is granted, for the sake of the argument, that terms such as 'water' have primary intensions that are available a priori, that will not suffice to support Jackson's form of argument.

Second point first: suppose 'water' has a primary intension—say X—and that it is knowable a priori that water has this primary intension. Consider the paradigm argument, put in terms of primary intensions:

(a) 60 percent of the globe is covered by H_2O.
(b) H_2O = the satisfier of X (the primary intension of 'water').
(c) Water = the satisfier of X.

Therefore,

(d) 60 percent of the globe is covered by water.

For the argument to succeed in showing that (d) is entailed on conceptual grounds by the microphysical facts, it is required that (c) be a conceptual truth, and that (b) be a microphysical truth. The first we are granting for the moment, but it should be immediately obvious that the apparatus of primary intensions does nothing to show the second. What (b) says is that the primary intension of 'water' maps the actual world onto exactly one item, H_2O. But we can conclude that this is a microphysical fact only if it is *assumed* that microphysical facts determine or include all the facts. The primary intension of 'water' is a function that takes W_{Earth} to water, $W_{\text{Twin Earth}}$ to twin water, etc. But we can't, without begging the question, take for granted that the *microphysical* description of W_{Earth} describes only W_{Earth}. Suppose there are two microphysically indiscernible possible worlds, W_{Earth} and $W_{\text{Super Twin Earth}}$. Suppose further that there are primary intensions that take W_{Earth} to H_2O, but $W_{\text{Super Twin Earth}}$ to something else. (Primary intensions are just functions, and given any difference in inputs, there will be some functions that yield a difference in outputs.) If this is true for the primary intension of 'water', then (b) will be a fact, but not a microphysical fact, or even supervenient on the microphysical facts. Unless we assume that the microphysical facts determine all the facts, or at least that the value of the primary intension for 'water' depends only on microphysical facts in the worlds that are

the arguments to the function, the argument won't work. If our candidate conceptual analysis were an explicit verbal definition, in the vocabulary of microphysics, then we could be sure that the analogue of premise (b) was a microphysical truth, but once we move to the more abstract and unconstrained representation of meaning given by the two-dimensional intensional framework, we can no longer assume that this premise is necessitated by microphysical theory, and even if we could assume this, it is not clear what it would mean to say that this premise is deducible a priori from microphysics. In any case, if we do allow ourselves to assume that it is a microphysical truth that H_2O is the satisfier of the primary intension of 'water', why do we not have equal justification to assume that it might be a microphysical truth about the actual world that pyramidal cell activity is the satisfier of the primary intension of 'consciousness'?

Earlier (in section 5) we discussed Jackson's remark that the medical profession would be outraged at the claim that they have not shown that fairies don't cause cancer. Our response was that although doctors can certainly rule out fairies as a cause of cancer, the claim that fairies don't cause cancer should not be regarded as literally a part of or deducible a priori from their theories. The discussion of the last paragraph shows that there is some value in translating the discussion into two-dimensional terms, since doing so reveals more starkly that the microphysical premise in Jackson's argument depends on the hidden assumption that the microphysical facts determine all the facts.

Let us now move to the other question: whether the two-dimensional apparatus provides a reason to believe that anything with the form of (c) is a conceptual truth. One might be tempted to think that the two-dimensional framework provides a recipe for determining the primary intension of any expression in the following way: Everyone can agree that *the world*—all that is the case—contains enough information to determine the semantic values (intensions of whatever kind, extensions, senses, contents, or whatever semantic values happen to be) of any expression that has a semantic value. So everyone can agree that if we are given an expression and a set of possible worlds in each of which the expression has some semantic value, this will determine a function taking the possible worlds into the semantic values, whatever they are, that the expression has in that world. If the values are secondary intensions,

then this function will be a two-dimensional intension that will determine a primary intension. But is this the *relevant* two-dimensional intension? We doubt that it is, since this function is not itself a kind of meaning that the expression has, but is a representation of the *possible* meanings that it might have. Primary intensions, understood this way, are like the primary weights of physical objects. Ordinary weight, which we might call 'secondary weight', is an empirical property of physical objects, but we can also define the primary weight of a thing in the following way: Primary weight is a property of a physical object that is knowable a priori: it is a function that takes any possible world and time into the secondary weight of the thing at that time. So, for example, you know the primary weight of Alexander the Great if you could infer his weight at any time in any possible world in which he exists from a complete description of that possible world. There is a world in which he is a giant of 500 pounds, and the value of the function for that world is 500 pounds. There is another in which he is a skinny 120 pounds, and the value for that world is 120 pounds. And so forth. Obviously, this is not something you need empirical information to be able to do.

Now consider the English word 'coumarone'. Do you know what it means?[11] Perhaps not, but if primary intensions are determined in the way we have suggested, then you know its primary intension. That is, if you were told enough about *the world,* or about any possible world, then without leaving your armchair you would be able to tell what that word meant and referred to in that world. Suppose that for all you know, our world might be one in which 'coumarone' refers to an extinct flightless bird. There is surely a possible world in which the word (or at least one orthographically like it) has such a meaning, and there is a two-dimensional intension that takes the actual world into the actual meaning of 'coumarone' and this counterfactual world into the meaning that it has there. Functions are cheap—there are lots of different two-dimensional intensions that could be defined—but is this the relevant one for that word?

One might reply that the relevant two-dimensional intensions are fixed by the dispositions that the speaker has in virtue of un-

[11] In case you are curious, it is another colorless liquid, not H_2O, but C_6H_4OCHCH. But of course it is not knowable a priori that that is what it is.

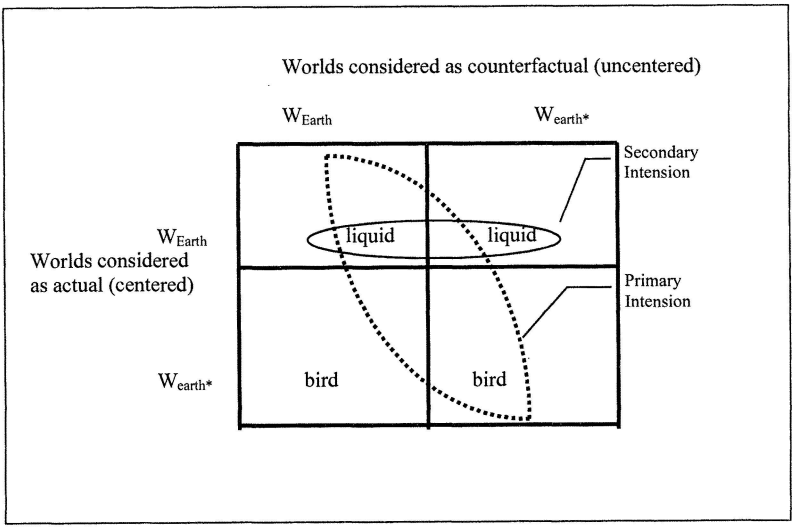

Fig. 2. 2-D intension of 'coumarone'

derstanding, or being a competent user of, the term in question, and so the reference of the word 'coumarone' in possible worlds which are compatible with the beliefs only of speakers who don't understand the word are not relevant. But this reply requires an account of how two-dimensional intensions are determined that is different from the one we are considering, an account that has not been provided. What exactly does one have to know to be a competent user of the term 'coumarone'? Is there any reason to believe that one can isolate a purely conceptual component of the knowledge of competent users of the term? The abstract two-dimensional framework itself is silent on these questions.

Alternatively, it may be suggested that the coumarone example is illicit because we are imagining speakers whose dispositions to use the word 'coumarone' are different from ours, or, better, for whom the functional role of 'coumarone' is different from ours. But this condition smuggles in a functionalist theory of concepts (via a functionalist theory of concept possession), and if we are prepared to buy that, it is unclear why we also need the primary intension theory. If the functionalist theory of concepts presupposes that there are analytic inferences that provide a purely conceptual component of the knowledge of competent users, that presupposition would have to be justified. If it does not suppose this,

we will have to deal with the familiar holism problem of nonanalytic functionalism. How, and in what ways, can people differ in roles for a word while still expressing the same concept by that word?

If the relevant two-dimensional intension is the one determined in the way we have been suggesting, then it may, in a trivial sense, be knowable a priori, but it won't do the work that needs to be done in the paradigm argument. Why not? The reason is that since it is not a kind of meaning, intuitions such as those appealed to in thought experiments about Jackson's Mary, or about Chalmers's zombies, are not relevant to the limits of this primary intension for 'consciousness'. No one will deny that the word 'consciousness', or at least a word that is phonologically and orthographically indistinguishable from it, might have referred to some brain process (or to a flightless bird, for that matter). But is a possible world in which 'consciousness' refers to a brain process a world in which the primary intension of *our* word 'consciousness' has a secondary intension that picks out a brain process, or is it a world in which the word 'consciousness' has a different primary intension from the one it has in our world?

The primary intensions defined by the procedure we have outlined are *derivative* from the actual and possible secondary intensions that a word has. To get the value of the primary intension at a given world, just find the secondary intension that the word has in that world, and the extension that it determines. But Chalmers insists that the primary intensions he is talking about are not derivative in this way. To determine the primary intension of some concept, it is just irrelevant what meaning the concept *would* have if used in some other possible world. It is *our* concept, as used in the actual world, that determines the referent in other possible worlds. "Given an individual's concept in thought, we can assign a primary intension corresponding to what it will pick out depending on how the actual world turns out" (Chalmers 1996, 65). "We can retain the concept from our own world, and consider how it applies to other worlds considered as actual" (366). Recall that, for Chalmers, a concept is something like a mental word—a syntactic or quasi-syntactic object. It is not a meaning, but something that has a meaning. But of course a word (or a concept, in this sense) will "pick out" or refer to something in a world only in virtue of its semantic properties—its meaning. So when we consid-

er how a concept from our own world applies in other worlds, we are considering the interpreted concept—the concept with its actual meaning. On the interpretation of primary intensions that Chalmers is rejecting, the meaning used to "pick out" an extension for the concept in a given possible world was the meaning that it had in that world. On Chalmers's intended interpretation, we "retain" the meaning the concept has in the actual world, asking what extension it determined in a given counterfactual possible world "considered as actual." But now we can see a dilemma: to get the primary intension, we carry our concept around from world to world, taking its actual meaning with it, to see what it picks out. In the case of our 'water' concept, the result is supposed to be that it picks out water on Earth, twin water on Twin Earth, etc. But what is it exactly that we carry, or "retain," from world to world? If it is the meaning in the ordinary sense (the "wide" meaning, or secondary intension), then it does *not* pick out twin water on Twin Earth, since as Putnam taught us 'water' refers to *water* (H_2O) everywhere (or to nothing in situations where there is no water). So it must be the narrow meaning—the purely conceptual content—that is retained. But the primary intension (or perhaps the two-dimensional intension), is supposed to be the explication of narrow, or purely conceptual, content. So this answer is of no help in explaining what the primary intension of a concept is. All that is being said is that the primary intension of a concept is the function that yields, for each possible world, the value of its primary intension for that world.

We can agree that *if* a word has a conceptual content that is available to anyone who understands the word, then this content might be usefully represented by a two-dimensional intension, but the apparatus by itself gives no support to the hypothesis that there is such a content. The apparatus *presupposes, rather than explains or justifies,* the distinctions that are required to factor the content of a concept into a priori and external components.

Chalmers does say some things about how to think about the primary intensions of our concepts, and so about what the facts are that give a concept its primary intension: we should reflect on what we should (not would, but would if rational) say if we were to find out certain things about the actual world. If we were to find out that the colorless odorless drinkable (more or less) stuff in rivers and lakes is XYZ, we would conclude that water is (necessar-

ily) XYZ. If we found that such stuff is not really a liquid, we would have found out, and we should say, that water is not a liquid. (But isn't it also true that if we were to learn that the word 'coumarone' referred to an extinct flightless bird, we would conclude that coumarones—as we would put it—are extinct flightless birds?) This seems to be armchair reasoning, reflection that does not include any obvious reference to real experiments, so it is tempting to conclude that this reflection just unfolds our concepts in a totally a priori way. But what this conclusion misses is that our reasoning about the proper epistemic response in various counterfactual situations is informed not only by our concepts, but by implicit and explicit theories and general methodological principles that we have absorbed through our scientific culture—by everything that the "we" who are performing these thought experiments believe. What people should rationally say in response to various hypothesized discoveries will vary depending on their experience, commitments and epistemic priorities.

We need not, however, put any weight on the claim that the methodological principles and priorities that we use to answer such questions are not a priori, or on any assumptions about what is or is not properly said to be part of conceptual content. The crucial question, for the issue we have been discussing in this paper, is whether a relevant contrast can be shown between the relation between water and H_2O on the one hand and the relation between consciousness and some brain process on the other.

Suppose, to try to get at the primary intension of the word 'consciousness', in the way suggested by Chalmers, we ask how "we" should respond if the neurophysiological and behavioral evidence were to provide dramatic support for the conclusion that there is (at least) a strong correlation between phenomenal consciousness and a certain very specific kind of brain process (for example, the kind suggested in the work of Crick and Koch discussed above). Further, suppose that the cluster of properties that a functionalist might be inclined to use to define phenomenal consciousness (the consciousness role) could be explained in terms of this brain process. A philosopher with physicalist leanings in such a possible world might reasonably conclude that these facts would justify identifying consciousness with the brain process. So an actual philosopher with such inclinations might reasonably conclude that the value of the function that is the primary intension of our word

'consciousness' for this possible scenario is this brain process, just as the value of the 'water' primary intension for our world is H_2O. Jackson and Chalmers would presumably disagree with this conclusion. Is this dispute a purely conceptual or semantic one, a dispute that shows that these different philosophers use the word 'consciousness' with different primary intensions? We doubt it, but however the disagreement is characterized, the issue does not seem to be different in principle from issues about the kind of scientific reduction and explanation that Levine, Jackson, and Chalmers are trying to contrast with the issue about the explanation of consciousness.

Consider once again an analogue of the paradigm argument for the case of a conscious state.

(a*) Pyramidal cell activity was rampant in medieval prisons.
(b*) Pyramidal cell activity = the satisfier of the primary intension of 'pain'.
(c*) Pain = the satisfier of the primary intension of 'pain'.

Therefore,

(d*) Pain was rampant in medieval prisons.

Our point is that there could be compelling motivation for (b*) and that (c*) has whatever a priori status (c) above has.

In the first nine sections of this paper, we discussed physicalism and reductive explanation in the context of putative explicit verbal analyses in microphysical terms of such ordinary concepts as *life* and *water*. We expressed skepticism about whether such concepts are a priori analyzable in microphysical terms (or whether there are microphysical sufficient conditions for their application). But Jackson and Chalmers are skeptical too, regarding every such analysis that they mention as only an approximation. To go beyond approximation, they recommend the two-dimensional framework that we have been discussing in this section. But as we have pointed out here, the two-dimensional apparatus does not in any way help to isolate an a priori conceptual component of content, but—at least as used by Jackson and Chalmers—merely presupposes that there is such a thing.

We have made three major points in this section

(1) We argued earlier that there is no reason to believe reduc-

tive explanation requires conceptual analyses of the sort "Water is the odorless, colorless liquid that. . . ." The move to primary intensions does not get around these considerations.

(2) The claim that H_2O is the (or even a) satisfier of the primary intension of 'water' is not a microphysical claim.

(3) The a priori status of the claim that water is the satisfier of the primary intension of 'water' does not escape the main criticisms applied to its less technical predecessor. And whatever a priori status it does have applies equally to the claim that pain is the satisfier of the primary intension of 'pain'.

New York University (Block)
Massachusetts Institute of Technology (Stalnaker)

References

Block, N. 1978. "Reductionism." In *Encyclopedia of Bioethics,* ed. Warren T. Reich, 1419–24. London: Macmillan.

———. 1991. "What Narrow Content is Not." In *Meaning in Mind: Fodor and His Critics,* ed. B. Loewer and G. Rey. Oxford: Blackwell.

Burge, T. 1979. "Individualism and the Mental." *Midwest Studies in Philosophy* 4:73–122.

Byrne, A. 1993. "The Emergent Mind." Ph.D. diss., Princeton University.

Chalmers, D. 1993. "Toward a Theory of Consciousness." Ph.D. diss., Indiana University.

———. 1996. *The Conscious Mind.* New York: Oxford University Press.

Crick, F. 1994. *The Astonishing Hypothesis.* New York: Scribner.

Crick, F., and C. Koch. 1990. "Towards a neurobiological theory of consciousness." *Seminars in the Neurosciences* 2:263–75.

Davies, M., and L. Humberstone. 1980. "Two Notions of Necessity." *Philosophical Studies* 38:1–30.

Davies, M., and G. Humphreys. 1993. *Consciousness: Psychological and Philosophical Essays.* Oxford: Blackwell.

Evans, G. 1982. *The Varieties of Reference.* Oxford: Oxford University Press.

Field, H. 1973. "Theory Change and the Indeterminacy of Reference." *Journal of Philosophy* 70:000–00.

Hill, C. 1997. "Imaginability, Conceivability, Possibility and the Mind-Body Problem." *Philosophical Studies* 87:61–85.

Horgan, T. 1984. "Supervenience and Cosmic Hermeneutics." *Southern Journal of Philosophy* 22 (supp.):19–38.

Jackson, F 1982. "Epiphenomenal Qualia." *Philosophical Quarterly* 32:127–36.

———. "Armchair Metaphysics." In *Philosophy in Mind,* ed. J. O'Leary-Hawthorne and M. Michael. Dordrecht: Kluwer.

———. 1994. "Finding the Mind in the Natural World." In *Philosophy and the Cognitive Sciences,* ed. R. Casati, B. Smith, G. White, 100–112. Vienna: Verlag Hölder-Pichler-Tempsky. Reprinted in *The Nature of Consciousness: Philosophical Debates,* ed. N. Block, O. Flanagan, and G. Guzeldere (Cambridge: MIT, 1997).

———. 1995. "Postscript to 'What Mary Didn't Know'." In *Contemporary Materialism,* ed. P. K. Moser and J. D. Trout. London: Routledge.

Kaplan, D. 1978. "Dthat." In *Pragmatics, Syntax and Semantics,* vol. 9: *Pragmatics,* ed. P. Cole. New York: Academic Press.

———. 1989. "Afterthoughts." In *Themes From Kaplan,* ed. J. Almog, J. Perry, and H. Wettstein, 565–614. Oxford: Oxford University Press.

Levine, J. 1983. "Materialism and qualia: the explanatory gap." *Pacific Philosophical Quarterly* 64:354–61.

———. 1993. "On leaving out what it is like." In Davies and Humphreys 1993.

Lewis, D. 1966. "An Argument for the Identity Theory." *Journal of Philosophy* 63:17–25.

———. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61:343–77.

Loar, B. 1990. "Phenomenal properties." In *Philosophical Perspectives: Action Theory and Philosophy of Mind,* ed. J. Tomberlin. Atascadero, Calif.: Ridgeview.

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

Nagel, T. 1974. "What is it like to be a bat?" *Philosophical Review* 83:435–50.

Segerberg, K. 1973. "Two-dimensional modal logic." *Journal of Philosophical Logic* 2:77–96.

Smart, J. J. C. 1959. "Sensations and Brain Processes." *Philosophical Review* 68:141–56.

Stalnaker, R. 1978. "Assertion." In *Syntax and Semantics,* vol. 9: *Pragmatics,* ed. P. Cole. New York: Academic Press.

———. 1990. "Narrow Content." In *Propositional Attitudes: The role of content in logic, language and mind,* ed. C. A. Anderson and J. Owens, 131–46. Stanford: CSLI.

White, S. 1982. "Partial Character and the Language of Thought." *Pacific Philosophical Quarterly* 63:347–65.