

*Explanatory Reduction, Conceptual Analysis,  
and Conceivability Arguments about the Mind*

BRIE GERTLER

University of Wisconsin, Madison

The current stand-off between reductionists and anti-reductionists about the mental has sparked a long-overdue reexamination of key issues in philosophical methodology.<sup>1</sup> The resulting debate promises to advance our understanding of how empirical discoveries bear on the numerous philosophical problems which involve the analysis or reduction of kinds. The parties to this debate disagree about how, and to what extent, conceptual facts contribute to justifying explanatory reductions.

My aim here is threefold: (a) to show that conceptual facts play a more significant role in justifying explanatory reductions than most of the disputants recognize,<sup>2</sup> (b) to furnish an account of that role, and (c) to trace the consequences of this account for conceivability arguments about the mind. I begin (Section I) by sketching an initial argument for the thesis that all justification for explanatory reductions is based in conceptual facts, in that our concept of a kind determines what qualifies as evidence for a reduction of the kind. The middle sections of the paper (Sections II–V) defend this thesis from recent influential objections. I extract from this defense a detailed model of how concepts contribute to explanatory reductions (Section VI). This model implies that reductionists cannot simply dismiss, as irrelevant, conceivability arguments against reductionism about the mind. In the final section (Section VII) I rehearse a familiar brand of conceivability argument, and sketch the reductionist strategies for defusing this argument which remain available on the model of explanatory reduction defended here. I then describe the anti-materialist rejoinders which that model makes available. I do not take a side in the debate over mental reductionism. My point is that the viability of reductionism must be decided on conceptual grounds and that, therefore, conceivability arguments are crucially important in evaluating materialism about the mind.

### I. The Argument from Relevance

A central assumption of the analytic tradition is this: analyzing a concept of a property or kind is a legitimate way to determine the essential nature of the property or kind. In fact, the tradition seems to presuppose that conceptual analysis provides substantial help in determining the essential nature of concrete phenomena such as pain and heat, as well as the nature of more abstract properties like justice and truth. But some philosophers now claim that the traditional reliance on conceptual analysis is misguided. These philosophers point to the surprisingness of scientific results, and argue that empirical discoveries can upset expectations based on deep-seated beliefs about the kind at issue, including beliefs that directly reflect our concept of the kind. Hence, empirical investigation can shed light on a kind independently of conceptual investigation. In the words of Block and Stalnaker, “we might have reason to believe that [an identity claim is true] even without the help of a conceptual analysis” (B&S 1999, 30). The proponents of this argument seek to exempt reductionist programs, such as reductionism about the mind, from conceptually-based objections.

In order to show that one *theory* (e.g., of pain or heat) is reducible to another, reductionists attempt to establish identities such as “pain = physical (or functional) state *c*” or “heat = molecular motion”. The identities themselves are called “explanatory reductions”, to mark that a term on one side of the identity sign belongs to a theory that is explanatorily more basic than the theory to which the other term belongs. Successful reductions increase a reducing theory’s explanatory power, for they expand the theory’s domain while retaining its simplicity. The question at issue, then, is this: how does the concept of a kind contribute to justifying a reduction of that kind?

The following is my argument to show that evidence for a reduction must be deemed as such by the concept of the reduced kind. I use, as schematic for explanatory reductions generally, “ $F = X$ ”, where “ $X$ ” belongs to a theory that is explanatorily more basic than the theory to which “ $F$ ” belongs. For instance, “ $F$ ” might refer to heat, and “ $X$ ” to molecular motion. A final preliminary note: given the way the argument employs “evidence”, there is no reason to place antecedent limits on what can qualify as evidence. So I do not exclude indirect evidence, reliabilist warrant, etc.

*The Argument from Relevance.*<sup>3</sup> For all subjects  $S$  and kinds  $F$  and  $X$ ,

1.  $S$  is justified in accepting “ $F = X$ ” **only if** there is an  $e$  (an actual event, fact, process, etc.) such that  $e$  warrants  $S$ ’s belief that “ $F = X$ ”.
2. For all  $e$ ,  $e$  warrants  $S$ ’s belief that “ $F = X$ ” **only if** “ $e$  qualifies as evidence for ‘ $F = X$ ’” is made true by  $S$ ’s concept [ $F$ ].
3. For all  $e$ , “ $e$  qualifies as evidence for ‘ $F = X$ ’” is made true by  $S$ ’s concept [ $F$ ] **only if** analysis of  $S$ ’s concept [ $F$ ] would reveal that  $e$  qualifies as evidence for “ $F = X$ ”.<sup>4</sup>

Therefore,

4. *S* is justified in accepting “ $F = X$ ” **only if** there is an *e* such that *e* warrants *S*’s belief that “ $F = X$ ” and analysis of *S*’s concept [*F*] would reveal that *e* qualifies as evidence for “ $F = X$ ”.

That is,

5. One is justified in accepting an explanatory reduction only if one’s evidence for the reduction would be deemed as evidence by an analysis of one’s concept. More generally, something can justify a reduction only if our concept of the reduced property or kind confers this justificatory status upon it.

The conclusion of this argument says that qualifying as evidence is ultimately a conceptual matter; I will refer to this conclusion as the “Conceptual Basis of Justification” (or “CBJ”) thesis.

The “Argument from Relevance” is a fitting name for this argument because the central premise, Premise 2, springs from the idea that purported evidence for a reduction is *relevant* to the explanandum only if one’s concept of the explanandum renders it relevant. Jackson also uses a relevance-based argument for the importance of conceptual analysis; he says, in a section provocatively entitled “The Case for Conceptual Analysis in a Sentence (or Two)”, “Only in that way [through conceptual analysis] do we define our subject as the subject we folk suppose is up for discussion”. (Jackson 1998, 42) (Jackson explicitly disavows a central consequence of my argument; see below, Section V, and my 1999b.)

Applying the Argument from Relevance to a familiar case (given in Putnam 1975) reveals its strong initial plausibility. Imagine that you discovered that the things you called “cat” were robots. Would this prompt you to accept eliminativism about cats? Or would it lead you to believe that “cat” refers to a robotic kind? By considering this question, you are undertaking conceptual analysis; this process reveals whether, given your concept of cats, the discovery would qualify as evidence that cats are robots. (Premise 3) Suppose that this process reveals that, on your concept [cat], the discovery would *not* qualify as evidence that cats are robots. Perhaps you consider *cat* an organic (non-robotic) natural kind. Then the discovery would not be relevant to the nature of the kind *cat*, but would instead support eliminativism about that kind. In other words, your concept [cat] determines what qualifies as evidence, for you, about *cats*. (Premise 2) But if you have no evidence that qualifies, for you, as evidence that “cat” is a robotic kind, then you are not justified in accepting an identity statement linking the kind *cat* to a robotic kind. (Premise 1) So conceptual analysis reveals what sorts of evidence could justify a given identity statement.

Of course, there are practical limits on conceptual analysis. First, undertaking conceptual analysis requires reflecting on a range of hypothetical scenar-

ios, and there is certainly no guarantee that someone who possesses a concept will be skilled at this. But concept possession does constrain one's dispositions to react to new discoveries, such as the discovery that the things we call "cat" are robots. (I return to this point in a moment.) Second, the results of conceptual analysis, regarding what qualifies as evidence for a reduction, are ordinarily quite general: for instance, we could not possibly complete the conceptual analysis which would be required to determine every potential discovery which could justify "cats =  $X$ s", for every  $X$  for which this statement could conceivably be justified. For this reason, "Evidence  $e$ " will typically denote evidence at a high level of generality, such as the target kind's basic functional structure, microphysical properties, or causal history. Third, conceptual analysis will not always yield determinate answers to the question whether a given discovery would justify a given reduction. This is as it should be, for it reflects the fact that many concepts are vague. Still, even extremely vague concepts are substantial enough to fix, at some level of generality, the discoveries which would justify reduction of the associated kind. My claim is that the contribution a particular piece of evidence makes to a reduction *derives from* such conceptual facts.

While I cannot give an exhaustive account of concepts here, a few remarks about their nature are in order. The Argument from Relevance employs "S's concept [ $F$ ]" rather than "*the* concept [ $F$ ]" or "*our* concept [ $F$ ]". I have no objection to the latter phrases, used to refer to that concept which competent English speakers associate with " $F$ ". But I am committed to the idea that shared concepts, or linguistic meanings of terms which express them, *derive from* individual concepts. Phrases like "the concept [water]" refer via generalizations from individual concepts, and thereby underwrite the meaning of "water". This individualism about concepts conflicts with the position of content externalists like Burge (1979). The scope of mental content is peripheral to the current project. Still, it will become clear that my account of how explanatory reductions are justified fits nicely into a larger individualistic picture.<sup>5</sup>

It is widely believed that possessing a concept requires having a particular set of dispositions. These may include the disposition to apply the associated term to certain things and not to others; the disposition to accept certain states of affairs as possible, and others as impossible; etc. This view, that having some such set of dispositions is necessary for possessing a concept, will be adequate for my purposes here. (I need not say what *suffices* for concept possession.) I remain neutral as to the precise nature of concepts, including whether possessing a concept explains, or rather reduces to, these dispositions.

Since beliefs are usually taken to be dispositional, taking some relevant set of dispositions to be necessary for concept possession accommodates the view that anyone who possesses a concept [ $F$ ] has a set of *a priori* beliefs concerning  $F$  or  $F$ s. Quine (1961) famously denies that there is a principled distinction between beliefs that define a concept and those that do not. Obviously, I

cannot address Quine's arguments here. But it is worth noting that, while the Argument from Relevance succeeds only if there are some conceptual truths, it doesn't simply presuppose that there are. It furnishes a reason to believe that there are, since it suggests that only conceptual truths can provide the link between the original target and a proposed reduction of it necessary to render the latter *relevant* to the former. This means that if there are any genuine reductions of original target properties, there must be some conceptual truths. (For recent defenses of the analytic / synthetic distinction from Quine's attack, see Sober and Hylton (2000) and Boghossian (1997).<sup>6</sup>)

Recent debate about the role of conceptual analysis has yielded four leading objections to the CBJ thesis. (For expository ease here and throughout, I assume the truth of "water = H<sub>2</sub>O", a relatively non-controversial example of an explanatory reduction.) (1) Water isn't grasped via a description. But conceptual analysis reveals only descriptive components of concepts. So conceptual analysis will not reveal what qualifies as evidence for "water = H<sub>2</sub>O"; at least some of the evidence for an explanatory reduction qualifies as such independently of (actual or potential) conceptual analyses. (2) The most plausible model of justification consistent with the CBJ thesis cannot ensure that H<sub>2</sub>O uniquely fits the conceptual bill, for [water]. So this model cannot explain how "water = H<sub>2</sub>O" could be justified. But "water = H<sub>2</sub>O" *is* justified. Hence, at least some of the evidence for an explanatory reduction qualifies as such independently of (actual or potential) conceptual analyses. (3) Identities are sufficiently justified by the explanatory force of the theories which they support; they do not require further, conceptually-based justification. So at least some of the evidence for an explanatory reduction qualifies as such independently of (actual or potential) conceptual analyses. (4) At most, conceptual analysis informs us about folk concepts; it doesn't tell us what the world is actually like. Since we are justified in believing that the world may not neatly match folk concepts, justification for an explanatory reduction need not be based in conceptual facts. Hence, at least some of the evidence for an explanatory reduction qualifies as such independently of (actual or potential) conceptual analyses.

I show that the Argument from Relevance withstands each of these four objections in the next four sections, respectively. My replies to these objections further elucidate *how* conceptual facts determine what qualifies as evidence for an explanatory reduction; I synthesize these results in Section VI. The final section applies these results to conceivability arguments about the mind.

## II. Objection #1

The first objection, made by Levine (1998) and (very briefly) by Block and Stalnaker (1999, p. 16), contends that subjectively inaccessible factors may fully justify some legitimate reductions. If there is an alternative means of qualifying as evidence for a reduction, a means independent of subjectively accessi-

ble factors, then a reduction may be justified independently of conceptual analysis, which after all reveals only subjectively accessible elements of concepts. This objection challenges the combination of Premises 2 and 3 of the Argument from Relevance, which together entail that only subjectively accessible elements of concepts can render a proposed reduction relevant to the target. It thereby threatens the basis for the CBJ thesis.

According to Levine's preferred alternative, the concepts of some properties concern their referents purely by virtue of an external—causal or nomic—link to them. That is, some properties have what Levine calls "non-ascriptive modes of presentation". Since I will borrow Levine's terminology here, I quote his explication of it at some length.

[L]et's distinguish two kinds of mode of presentation (where by a 'mode of presentation' is meant the means by which a representation connects to its referent): ascriptive and non-ascriptive. An ascriptive mode is one that involves the ascription of properties to the referent, and it's (at least partly) by virtue of its instantiation of these properties that the object (or property) is the referent. A non-ascriptive mode is one that reaches its target, establishes a referential relation, by some other method. The object isn't referred to by virtue of its satisfaction of any conditions explicitly represented in the mode of presentation, but rather by its standing in some particular relation [e.g., a causal or nomic relation] to the representation. (1998, 457)

[Non-ascriptive modes] establish relations... "behind the scenes", not by being cognitively grasped by the subject. The subject's competence with the term, her "knowledge" of the meaning, consists entirely in her instantiating the requisite relation to something in the world. (ibid., 458)

Levine uses the claim that some properties (including, on his view, qualia properties) have purely non-ascriptive modes of presentation to discount the force of conceivability intuitions about those properties. If a property has a purely non-ascriptive mode of presentation, then conceptual analysis, which uncovers only ascriptive aspects of the property's mode of presentation, will not elucidate the property. Assuming that at least some properties with purely non-ascriptive modes are reducible, the justification for such reductions is wholly independent of ascriptive conceptual facts, and so the CBJ thesis is false.

My response to this first objection is to deny that any properties have *purely* non-ascriptive modes of presentation. (As Levine defines these terms, all non-ascriptive modes are purely non-ascriptive.) The rest of this section will be devoted to establishing this point. I begin with the parallel linguistic point: I argue that Kripke's arguments against the descriptive theory of reference, which (together with Putnam's closely related arguments) are perhaps the foremost source of reservations about conceptual analysis, suggest that ascriptive ele-

ments of the concept associated with a linguistic term *render* non-ascriptive (causal / historical / nomic) factors relevant to determining its reference. I then make a parallel case regarding properties: some properties have partially causal or nomic modes of presentation, but those properties' modes have non-ascriptive elements in virtue of their *ascriptive* modes.

Kripke argues against the descriptive theory of reference, according to which the referent of a term is whatever fits (most of) the descriptions we associate with the term, by charging that it can land on the wrong referents. On his view, the descriptive theory overlooks the referential work done by factors external to the subject, including historical and environmental facts. For instance, our concept [gold] is indexed to certain physical facts which—perhaps together with expert uses of “gold”—partially explain how our term “gold” refers to actual *gold*. Kripke's argument rests on conceivability intuitions, e.g., it is conceivable that some of the yellow, coveted metal around here is not gold; it is conceivable that there is something else (e.g., iron pyrites) which fits more of the descriptions we associate with “gold” than gold itself does; it is conceivable that we discover that gold has atomic number 37; but it is inconceivable that, while all the gold around here has atomic number 79, gold on Twin Earth has atomic number 37. Such conceivability intuitions show that ascriptive features don't exhaustively determine the referent of “gold”.

Some take Kripke's arguments to show that descriptive theories err in overestimating the importance of our concepts. But I take them to show that descriptive theories err in *misconstruing* our concepts, by identifying them with a cluster of pure descriptions. They show that some concepts are *indexed to* non-ascriptive (external) factors. These arguments depend on our concepts having some ascriptive components, in the sense that subjects who possess the concept, and who understand the associated term, are disposed to recognize which sorts of factors secure that term's reference. Moreover, these ascriptive components—our dispositions to identify particular referents in particular hypothetical cases—render non-ascriptive factors relevant to determining reference. For Kripke's hypothetical scenarios reveal how we conceptualize the referent of “gold”—as a kind which in fact has certain superficial features (yellowness, hardness) which we use to pick it out, but which is individuated by another, non-superficial property (its atomic weight).

The observation that modes of presentation often differ from individuation conditions bolsters one of Kripke's central claims, namely, that some *a priori* truths are contingent. E.g., one could conceptualize gold in a way that makes the following *a priori* true: “Gold is the substance instantiated by the items over there, or at any rate, by almost all of them” (Kripke 1972, 135). In this case, “the substance instantiated by the items over there, or at any rate, by almost all of them” serves as gold's mode of presentation for the subject. The fact that this is gold's mode of presentation can be known through conceptual reflection, and is therefore *a priori*. But Kripke denies that all *a priori* truths are necessary. First, gold could have had a different mode of presentation.

Second, and more importantly, it is not a necessary truth that gold has the property actually serving as its mode of presentation. It is contingent that there is gold over there, and hence contingent that gold has the property which serves as its mode of presentation.

What's crucial to the present point is that the subject can know, through reflection alone, how "the substance instantiated by the items over there, or at any rate, by almost all of them" fixes the reference of her term "gold". Particular non-ascriptive facts (e.g., that the items over there have atomic number 79) are relevant to reference only insofar as ascriptive factors render them relevant. In this case, the ascriptive factors include gold's mode of presentation, for the subject, and the fact that this mode differs from the condition she takes to individuate gold. As before, these factors are manifested in the subject's disposition to apply "gold" in a range of situations. So non-ascriptive factors govern reference only because, and to the extent that, ascriptive factors—how we conceptualize the referent—confer this governing role upon them.

I honestly don't know whether this construal of the argument is that which Kripke intended. But I do think that Kripke's argument succeeds in establishing his anti-descriptivist conclusion only on this construal.<sup>7</sup> Against my position here, one could claim that understanding a term doesn't require even an implicit grasp of how the term's reference is fixed. In that case, one who understands a term needn't know what sorts of factors fix reference, and so these factors could be purely non-ascriptive. But this alternative interpretation leaves us with no reason to accept that non-ascriptive factors play any role in securing reference. Kripke's argument crucially depends on the conceivability of various hypothetical scenarios, and the associated dispositions to identify particular referents in counterfactual situations. Suppose there is an intransigent descriptivist who is disposed to accept the following: gold around here has atomic number 79, and chemists individuate substances by atomic number; but if there is stuff on Twin Earth which looks like gold then that stuff *is* gold, even if it has atomic number 17. If most of us had similar descriptivist dispositions, then Kripke's anti-descriptivist conclusion about "gold" would be false. If only a few occupy such a stance, then I think the right thing to say is that those people don't understand "gold" as the rest of us do.

Of course, someone who defends a false theory of reference can nonetheless be a competent speaker, for one's verbal dispositions may be at odds with one's professed theory. Reading Kripke's argument led many philosophers to recognize that their theory of reference was at odds with their verbal dispositions, and hence to revise their theory of reference. It did not lead them to alter their referential *practices*. In any case, Kripke's argument against descriptivism depends on our actual disposition to use "gold" in a way inconsistent with descriptivism. Kripke has not shown that causal links can secure reference without being cast in this reference-securing role by conceptual, ascriptive factors.

Let us turn from linguistic entities to our central concern, concepts of natural kinds like *water*. To determine the non-ascriptive elements in water's mode

of presentation, we ask questions such as, “would a given stuff be water, if it had a different microstructure than the actual watery stuff around here?” Perhaps grasping [water] doesn’t require understanding that microstructure is crucial. But some understanding of what is crucial must be in place, in order for the subject to truly have the concept [water]: perhaps it is enough that one is disposed to treat, as satisfying [water], whatever it is that scientists around here find crucial to the watery stuff around here.

Even if the subject has only a tacit, minimal conception of water, her judgment employing [water] is irrelevant to *water* unless she can recognize, at some level of generality, which kinds of conditions an object must fulfill in order to satisfy her concept. And she is unjustified in accepting a proposed reduction of water as a reduction of that which she conceptualizes as [water], unless she has evidence for the reduction which qualifies as evidence according to her concept. Without some link between the purported evidence for a reduction of water and one’s concept [water], there are no grounds for accepting the evidence as relevant to *water*. So it is water’s *ascriptive* mode of presentation which dictates that instances of water (items which satisfy [water]) must have the correct causal connection to scientific authorities, antecedent usages and/or microphysical facts. That is, it is ascriptive factors which index [water] to the relevant empirical, non-ascriptive factors.

Competent speakers and concept possessors are able to determine, within limits, under what conditions a statement concerns water, and under what conditions a putative “water” statement changes the topic from water to something else. The ability to determine relevance is anchored in the mastery of concepts and associated terms. This mastery doesn’t require an ability to exhaustively determine referents—after all, it is *conceivable* that XYZ is in the extension of [water] or “water”. Nor does it require an ability to articulate the kinds of conditions which secure reference—few who grasp [water] or “water” are prepared to report that something qualifies as water just in case it has the same microstructure as the actual watery stuff around here. But facility with a concept or term does require an ability to *recognize* the kinds of conditions which secure reference. That is, it requires that one be generally disposed, within the qualifications registered in the previous section, to make appropriate judgments about referential relations in counterfactual situations.

For a final example, consider Burge’s Oscar, a character with an exceedingly weak descriptive grasp of [arthritis]. (Burge 1979) When Oscar reports to his doctor that he fears he has arthritis in his thigh, the doctor tells him that arthritis is a joint disease and so cannot afflict the thigh. Oscar’s acceptance of the doctor’s response shows that he is disposed to believe, in the face of that testimony, that arthritis cannot afflict the thigh. Moreover, Oscar presumably has other key dispositions regarding [arthritis], e.g. to deny that “I have arthritis” can be made true by pure stipulation on his part. And so even for Oscar, arthritis has an ascriptive mode of presentation: [arthritis] is such that its extension is fixed by *these* sorts of empirical facts and/or social practices,

and not by *those*. Oscar's concept of arthritis explains or consists in his disposition to apply [arthritis] consistent with its extension being fixed in that way.<sup>8</sup>

Let us review. Objection #1 is that, contra the CBJ thesis, reductions are sometimes justified through non-ascriptive facts alone. I have argued that the particular non-ascriptive elements in a property's mode are determined by the particular *ascriptive* elements in the mode. So the ascriptive aspects of a property's mode of presentation, those in-principle accessible through conceptual analysis, are what give non-ascriptive aspects their status as evidence. Since non-ascriptive factors qualify as evidence only in virtue of ascriptive factors, dispensing with ascriptive factors would undercut the support for causal theories of reference. (This consideration does the primary work in Bealer's 1987 paper.<sup>9</sup>)

The arguments for the causal theory of reference turn on the claim that, to use Putnam's phrase, actual language practices exhibit a "division of linguistic labor" between ordinary folk and external authorities. Because the division of linguistic (and conceptual) labor is important throughout the paper, I want to end this section by drawing out a pertinent consequence of the current discussion. To express the relation between the folk and external authorities, apropos of linguistic and conceptual labor, I prefer the term "deference", which underscores the above-described role of folk concepts in determining *how* labor is divided, to the more passive "division of labor". Deference is willingness to accept the verdict of an external authority about the essence of a kind, and hence about the extension of a term. In this context, "authority" (and, correspondingly, "deference") has an unusually broad application: it can refer to empirical facts, such as the fact that the watery stuff around here has microstructure H<sub>2</sub>O, as well as to individual humans and scientific communities. For instance, my disposition to deny that water *could* (metaphysically) be XYZ reflects my deference, regarding water, to particular microphysical facts; Oscar's disposition to deny that his former self-diagnosis ("I have arthritis in my thigh") even *could* be true reflects his deference, regarding arthritis, to medical authorities.

Scientific experts also exercise deference. They typically defer to certain facts as determining the essence, and hence the extension, of kinds they study. Deference to microphysical facts explains the chemist's discovery that jade isn't a (unitary) natural kind; deference to causal-historical facts explains the biologist's discovery that frogs and toads belong to the same genus. Is all deference to experts, then, ultimately deference to facts? Perhaps. I shall leave this issue open, as nothing here rests upon it.

To modify Putnam's metaphor, folk are not only laborers in the factories of reference; they are also the factory managers, who determine how the labor is to be divided. It is not that each of us *deliberately chooses* to defer to authorities about the essence of water, of course. We acquiesce in this sort of practice.<sup>10</sup> Still, human or factual authorities can justify proposed identity statements which draw on our concepts only because, and to the extent that, folk implic-

itly assign certain kinds of referential labor to them. External factors cannot wrest final control over reference from the conceptual managers, ordinary folk who bestow authority upon them through deference. With some hesitation, I dub this view *Reference Deference*.<sup>11</sup>

### III. Objection #2

Block and Stalnaker (“B&S” hereafter) offer a detailed argument to show that there is no conceptual analysis of [water] which, in conjunction with micro-physical facts, suffices to justify “water = H<sub>2</sub>O”. On their view, conceptual factors do not prescribe the role which non-conceptual factors play in justifying reductions. This implies that not all evidence for explanatory reductions owes its status as evidence to our concepts. It thereby implies that Premise 2 of the Argument from Relevance and the CBJ thesis are both false.

B&S target the *two-dimensional account* of the contribution conceptual analysis makes to explanatory reductions. The two-dimensional account has been developed and defended by friends of conceptual analysis like Jackson (1998) and Chalmers (1996), as a way to reconcile the *a posteriori* status of some identities with *a priori* philosophical methodology.<sup>12</sup> It provides the most plausible explanation of the vital contribution conceptual analysis makes to justifying reductions, and thereby supports the CBJ thesis. And as we will see, it meshes well with (though it does not entail) Reference Deference. If the two-dimensional account does not explain how reductions are justified, my central claims are in jeopardy.

The two-dimensional account envisions explanatory reduction as the result of two steps. Step 1 is conceptual analysis. In the case of water, Step 1 yields a definite description indexed to the surroundings and rigidified: e.g., water = that which actually plays the “water” role, that is, the actual watery stuff around here. Step 2 involves determining, usually through empirical investigation, what meets the indexed description yielded by the first step. For instance, H<sub>2</sub>O is what satisfies the description “the actual watery stuff around here”. Not just any specification of what satisfies “the actual watery stuff around here” will do, of course. B&S assume that the two-dimensional account of explanatory reduction is committed to Step 2 specifying a microphysical structure; they therefore call Step 2 “the microphysical premise”. (I question this assumption below.)

B&S argue that no two-dimensional account, based on conceptual analysis and microphysical research, can guarantee that there is a unique thing which plays the water role. Without the guarantee of uniqueness, a proposed identity (such as “water = H<sub>2</sub>O”) is threatened, for if there is something other than H<sub>2</sub>O which also plays the water role, that other thing has an equal claim to identity with water. Identity is, of course, transitive and symmetrical; water cannot be identical to two distinct things. This is the *Uniqueness Problem* for the two-dimensional account.

The Uniqueness Problem captures a central motivation for doubting the importance of conceptual analysis, namely, conceptual modesty. It seems a sort of arrogance to suppose that our concepts should somehow circumscribe empirical discoveries. The Uniqueness Problem points to limitations in our concepts, suggesting that empirical discoveries can legitimately outstrip them, contra the Argument from Relevance.

I shall defend the two-dimensional account from the Uniqueness Problem. I address, in turn, each of the Uniqueness Problem's two components, as identified by B&S: (i) in addition to  $H_2O$ , there may be something *nonphysical* which plays the definitive water role, that is, which fits the description "the actual watery stuff around here"; and (ii) there may be more than one *physical* thing which plays this role.

(i)

B&S observe that it is beyond the province of microphysics to deny that anything nonphysical plays the definitive *water* role, in addition to  $H_2O$ .<sup>13</sup> This limits the possible force of the so-called microphysical premise (Step 2); it may state that  $H_2O$  plays the water role, but it cannot assert that  $H_2O$  *uniquely* plays this role. What, then, safeguards "water =  $H_2O$ " from the threat that something nonphysical may have equal claim to be (identical to) water?

The obvious answer is that this safeguard derives from Step 1 of the account, namely, the analysis of [water]. Specifically, our concept of water might exclude the possibility of nonphysical water; it may be a conceptual truth that water is physical. B&S allow that this may be true of water, but they deny that it can be generalized. "[F]or 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." (B&S 1999, 18). In other words, some physical things are conceivably nonphysical, yet are physically reducible. Even if water is not among those things, they argue, that there are such things shows that the two-dimensional account provides, at best, an incomplete picture of reductive explanation.

That there are physical things which are conceivably nonphysical fails, in fact, to show that the two-dimensional account is incomplete. Recall that on my view deference is necessary to link kind terms such as "water" to microphysical facts in the speaker's environment. If we defer directly to microphysical facts regarding a particular kind, then the kind is *not* conceivably nonphysical and hence the "nonphysical competitors" threat is empty. Now suppose instead that we defer directly to scientific authorities, and only indirectly to microphysical facts. I shall argue as follows. Deference to scientists either reflects a conceptual truth, that the kind is a physical kind, or it does not. If it does reflect this, then the kind is not conceivably nonphysical and the "non-physical competitors" threat is once again empty. (Of course, the "physical competitors" threat still stands; this is discussed in (ii), below.) Alternatively, if

deference to scientists does *not* rule out the conceptual possibility of nonphysical water, then the two-dimensional account can, at least in principle, ensure uniqueness. It will no longer be a “conceptual plus (specifically) microphysical facts” account, but my fundamental point, the CBJ thesis, will remain untouched.

It is plausible that we defer to scientists *qua* expert practitioners of a broadly empirical, scientific methodology, and that deference regarding ontology is a consequence of this. Suppose, for the moment, that deference to the results of a scientific methodology generally reflects an implicit commitment to physicalism about the kind in question. Then we would not defer to empirical scientists regarding physical things which are conceivably nonphysical. In other words, on the supposition that scientific authorities are seen as authorities about the physical, deference to empirical scientists would mean that the kind in question is *not* conceivably nonphysical. Deference would thus defuse the “nonphysical competitors” component of the Uniqueness Problem. Still supposing that whatever is thought to be discoverable by scientific authorities is thereby thought to be physical, a kind which *is* conceivably nonphysical would not be the subject of deference, at least not deference to those scientists.<sup>14</sup> If no deference is at work, the two-dimensional account is unavailable. This is not, however, a problem for the friends of conceptual analysis, since the two-dimensional account is intended to explain how the overarching importance of conceptual facts is compatible with the reference-securing role of non-ascriptive (external) factors. It achieves this by showing that the contribution of external authorities (Step 2) takes place only within the limitations set by our concepts (Step 1). If there are kinds about which we do not defer, then no such account is necessary, for in such cases our concepts would perform all of the work necessary for reference and for justifying reductions. So, on the supposition that ontological (physicalist) commitments directly follow from methodological (empirical) commitments, the possibility of nonphysical competitors poses no difficulty for the CBJ thesis.

Now suppose, instead, that the contrary is true: physicalism about a kind does *not* follow automatically from deference to empirical scientists about that kind. That is, suppose the folk do not believe that only physical kinds are the proper targets of empirical science. Then the conceptual analysis premise does not exclude nonphysical competitors. Rather, these fall within the scope of Step 2 of the two-dimensional account (which need not, then, always be a “microphysical” premise). In other words, if nonphysical competitors are conceivable, then our current supposition relegates the task of sorting out such competitors to empirical science, which is responsible for specifying what it is that satisfies the analysis yielded by Step 1. Empirical scientists are then in a position to determine whether the candidate at issue is unique, that is, to determine whether it faces physical or nonphysical competitors.

What if we think that water is probably physical, but we find it (barely) conceivable that water is nonphysical? In that case, we might accept a physi-

calist reduction of water based on scientific evidence, even if we deny that scientists investigate the nonphysical. We'd thereby allow that scientists can determine whether a physical candidate is explanatory enough to justify (together with our intuition that water is a non-disjunctive kind) accepting that water is in fact physical. As with the previous results, this depends on the nature of our concept [water].

B&S use the Uniqueness Problem to refute a particular version of the two-dimensional account, according to which Step 2 always identifies a microphysical property. This seems to me a needless restriction, given their larger aim: to contest the claim that "conceptual analysis is *necessary* to close the explanatory gap", by showing that we can justify identity claims "even without the help of conceptual analysis". (B&S 1999, 2;30) In any case, the CBJ thesis is not committed to the claim that conceptual facts allow only microphysical facts to qualify as evidence for a reduction. (See note 14.) It makes the more general claim that any sort of fact contributes to justifying a reduction only if it qualifies as evidence by virtue of conceptual facts.<sup>15</sup>

If we defer to empirical scientists, regarding water, then it is a *conceptual* truth that the essence of water is empirically discoverable. In that case, Step 2 can rule out all relevant competitors, physical and nonphysical. If we do not defer to scientists or other authorities, regarding water, then friends of conceptual analysis (and of the CBJ thesis) need not concern themselves with accommodating Step 2. This brings us to the second component of the Uniqueness Problem.

(ii)

The second component of B&S's Uniqueness Problem concerns the possibility that more than one physical thing fills a particular role. B&S note that there are three ways we could react to a discovery that more than one physical thing fills a particular role: adopt eliminativism about the type in question; construe the type as a disjunctive non-kind; or take the type to be a "superficial role property" type rather than a "role filler" type.

B&S acknowledge that the first, eliminativist possibility poses no threat to the two-dimensional account of explanatory reduction, since no explanatory reduction occurs in that case. The two-dimensional account is also consistent with the second, disjunctive possibility. For while it is not eliminativist about the role filler, the disjunctive option denies that what fills the role is a genuine *natural kind*, and so there is no explanatory reduction there, either.

It is the third possibility, according to which the term in question names a role property, which is alleged by B&S to pose a problem for the two-dimensional account. They offer the following example. "If... 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." (B&S 1999, 22) They are not claiming that this is the best construal of "jade", but just that this type of construal may be appropriate in some cases. This third

possibility is purportedly non-eliminativist, and it construes the kind in question as a *natural* kind. Given that the two-dimensional account identifies role fillers, rather than role properties, this option violates that approach. Since this option may be valid in particular cases, B&S conclude, the two-dimensional account cannot account for some justified explanatory reductions.

Allow that the role property option is valid in some cases. To evaluate them, we must ask: does conceptual analysis (Step 1) reveal that the concept at issue is a “role property” concept? If the answer to this question is “yes”, then this is not an objection to the (unrestricted) two-dimensional account or to the CBJ thesis. For in that case evidence about the role property will qualify as evidence for a reduction according to our concept. Alternatively, if conceptual analysis reveals that the concept is a “role filler” concept, then the “role property” construal does not reduce the kind originally at issue. Rather, that construal is subtly eliminativist, though not about whether there is a natural kind which answers to the *term* “jade”. It is eliminativist about the existence of a natural kind which answers to our *original concept* [jade], according to which jade things share a (probably microphysical) property which explains but does not reduce to the superficial features listed above. So if our concept is a “role filler” concept, the “role property” construal introduces a new explanandum. (The term “jade” may yet express the superficial role property, for concept individuation may not neatly follow word individuation.)

This is not, then, a case of explanatory reduction without conceptual analysis. So this third option either yields a reduction which is sanctioned by our concepts or supports eliminativism about a natural kind answering to our original concept [jade]. It therefore fails as a counter-example to the claim that qualifying as evidence for an explanatory reduction is a conceptual matter.

All three possible reactions to discovering that more than one physical thing fills a definitive role thus qualify as eliminativist, but each is eliminativist about different targets. The first is eliminativist about anything named by, say, “water”. The “disjunctive non-kind” possibility allows that something answers to the “water” category; its eliminativism lies in the denial that this names a *natural* kind. The “role property” possibility allows that there is a natural kind answering to the term at issue, but maintains that the reduction either fits the two-dimensional account (and the CBJ thesis) or is eliminativist about the referent of “water” in that term’s original sense.

I have responded to Objection #2, the Uniqueness Problem, as follows. The conceivability of nonphysical things which play the role definitive of water does not show that any evidence for “water = H<sub>2</sub>O” qualifies as such on non-conceptual grounds. If we can defer to scientists about something without thereby being committed to its being physical, then it is within the province of empirical science to determine whether there are nonphysical role fillers. Alternatively, if our deference to empirical scientists commits us to physicalism about the kind in question, then we do not defer regarding anything conceivably nonphysical. The possibility that more than one physical (empirically

discoverable) thing could play a given role also fails to show that there are grounds for explanatory reductions independent of conceptual facts, since this result warrants a reduction only if the original concept is not a role filler concept. If the original concept *is* a role filler concept, then this result warrants eliminativism about the original explanandum.

#### **IV. Objection #3**

The third objection to the CBJ thesis states that actual scientific reductions do not always need conceptual support, since some of these reductions are fully justified by the explanatory force of theories to which they contribute. Levine expresses this objection as follows: “That water is H<sub>2</sub>O is not the conclusion of any derivation. Rather, it functions as a premise in various explanatory arguments which have descriptions of water’s macro properties as their conclusions. When asked for the justification of the premise itself, the answer is that it’s justified because of the explanatory role it plays.” (1998, 462) B&S take a similar line. They note that identities serve as explanatory bedrock; while mere correlations stand in need of explanation, identities do not. Postulating identities increases simplicity, and hence explanatory force, by doing away with what Smart (1959) called “nomological danglers”.<sup>16</sup> The suggestion is that, if a proposed reduction contributes to the explanatory power of a theory, we need not evaluate the proposed reduction by its loyalty to our concepts. Explanatory force can thus partially justify a reduction independently of (actual or potential) conceptual analyses.

I argued above that the causal theory of reference is not an independent alternative to theories which trade on ascriptive elements of concepts or terms. For similar reasons, explanatory considerations are not an independent alternative to conceptual considerations, as a source of justification. The fact that a proposed reduction would increase a theory’s explanatory force qualifies as evidence for the reduction only if our concept of the target includes an appropriate deferential component. We might well be disposed to accept the increase in explanatory force as evidence for a reduction. But this disposition reflects our concept of the target: it is a conceptual truth that water is a natural kind, individuated by whatever it is that explains the macro properties of the watery stuff around here. What makes for explanatory power is itself something about which we likely defer. Still, the explanatory power of an identity statement justifies it only insofar as we are disposed to treat explanatory power as an authoritative factor. And we are thus disposed only by virtue of deference to explanatory power. This deference is usually indirect, in that we defer to experts who evaluate the contribution a proposed reduction would make to the explanatory power of a theory. In any case, the core point still stands: an identity statement can reduce a kind only if it concerns that kind, and it is conceptual facts about the kind which determine what sorts of statements are relevant to it.

Of course, some concepts have relatively insignificant *descriptive* components, and hence place only minimal *descriptive* constraints on candidate reductions. The extent of deference, regarding the extension of a given concept, determines the influence of non-descriptive considerations such as explanatory power. My concept [quark] is minimally descriptive, in that it is nearly exhausted by the content “whatever expert physicists actually mean by ‘quark’”.

While some concepts have insignificant descriptive components, there are arguably others about which we defer very little or not at all. The most likely candidates for these are abstract concepts, such as [seven] and [justice], and—as Kripke argued—phenomenal concepts such as [pain]. Simplicity considerations may yet influence explanatory reductions of these things; but, to the extent that we retain authority about these things, the simplicity at issue is conceptual rather than empirical. And it can happen that considerations of conceptual simplicity result in disjunctive analyses of what was allegedly a single concept. When this occurs, it is claimed that the putative concept is not unitary but instead runs together two distinct concepts. Tellingly, papers which attempt to reveal such conflation often have titles like “Two Concepts of Liberty” (Berlin 1969) and “Two Concepts of Consciousness” (Rosenthal 1986).

Explanatory considerations do substantially shape those reductions which concern concepts with a relatively small descriptive component. But they do so only in virtue of our dispositions to see them as decisive, and these dispositions reflect our concept of the target. So explanatory force does not constitute an independent alternative, to conceptual analysis, as a source of justification. The CBJ thesis withstands this third objection.

## V. Objection #4

Surprisingly, the fourth objection is voiced by Jackson, an outspoken proponent of conceptual analysis. Jackson expresses hesitation about conceptual analysis “being given a major role in an argument concerning what the world is like.” (1998, 43) He credits this hesitation with his rejection of his own previous (1982) anti-materialist conceivability argument. (1998, 43, note 21) At most, he says, conceivability arguments can show whether a proposed reduction fits the folk concept of the entity in question. But the folk could be wrong about “what the world is like”, and so conceivability arguments do not reveal what the world is like.

I find Jackson’s view puzzling, given that his overall approach exploits the intuitions behind the Argument from Relevance. If violating the results of conceptual analysis means changing the topic, how can a proposed reduction which clashes with the folk concept succeed? Such a proposal would seem doomed to eliminativism. The only grounds I can see for accepting a reduction which violates folk concepts, as relevant to the property or kind in question, are those which the above objections employ. But we have seen that the CBJ thesis survives these objections. Of course, even if much of the watery stuff around here

proved not to be  $H_2O$ ,  $H_2O$  could still be (identical to) *something* of interest. The Argument from Relevance, together with the two-dimensional account, entail only that, if enough of that stuff isn't  $H_2O$ , then  $H_2O$  isn't *water*. Similarly, if *c-fibers firing* fails a conceptual test for pain, then while the firing of *c-fibers* may be of interest, and may stand in some significant relation to pain, pain isn't identical to the firing of *c-fibers*.

The folk are not generally authoritative on whether there is anything which fits their concept; that is why there is logical space for eliminativism about folk kinds, e.g., ghosts.<sup>17</sup> And satisfying the two-dimensional account does not insulate a reduction from skeptical worries. Perhaps the actual watery stuff around here isn't  $H_2O$  at all, and our (apparent) evidence for Step 2 stems from a vast scientific conspiracy or massive hallucination. Or perhaps our evidence for Step 1 is flawed, in that our reactions to imagined hypothetical scenarios are not a reliable guide to our actual dispositions. Yet these skeptical possibilities don't create a gap between what folk concepts justify and what is actually justified; they provide equal grounds for doubting " $water = H_2O$ ". Jackson is right that the world may not neatly fit folk concepts, but this doesn't provide a reason to downplay their importance. For they offer our ultimate purchase on the way the world is. The fact that the world may not neatly fit folk concepts shows merely that the reductions we endorse may be false, i.e., that our methods for establishing identities don't generally yield certainty.

Instead of limiting the role of folk concepts, a better strategy for Jackson would be to claim that every folk concept includes a sort of *global* deference to external authorities. On that view, the folk maintain a blanket fallibilism about the world, so that any particular belief (including allegedly "conceptual" beliefs) could be shown false by some external authority or other. This would reconcile the Argument from Relevance with the claim that folk concepts might not reflect the way the world is. This view is consistent with CBJ, since conceptual analysis is required to reveal the global deference and hence to show that a proposed reduction is relevant to the explanandum. Unlike Jackson's stated view, this position would deny that the folk could be *ultimately* mistaken about "what the world is like"; ultimate error—such as that which Jackson now attributes to dualists—would be precluded by the limit which blanket fallibilism places on folk commitments. But perhaps Jackson would accept this consequence, since justified reductions would avoid violating folk concepts only trivially. That is, it allows that the folk are ultimately *ignorant* about "what the world is like", even if their blanket fallibilism rescued them from (positive) error.

I reject the position just described, as an inaccurate portrayal of folk concepts. To defend it one must show, against strong intuitive arguments (including Kripke's 1972, Lecture III) that even phenomenal concepts like [pain] contain an element of deference. (I return to the issue of deference and phenomenal concepts briefly in the final section.) And while it is strictly consistent with the Argument from Relevance and the CBJ thesis, it violates the spirit

of those claims by severely deflating the constraints imposed by our concepts. Finally, I doubt that this blanket fallibilism allows for broadly Fregean individuation conditions for concepts, since it allows the fallibilism to override particular ascriptive conditions which individuate concepts. In any case, a broad worry about the trustworthiness of folk concepts doesn't provide a reason, beyond those already discussed, to reject the CBJ thesis.

## VI. Synthesis of the Above Results

The CBJ thesis says that what qualifies as justification for a reduction is, without exception, an ultimately conceptual matter. The foregoing discussion illustrates *how* conceptual facts underlie justification for reductions. Section II showed that our concept of a kind dictates what sorts of facts, or fact-finding methods, qualify as authoritative about the kind. Applied to the two-dimensional account of explanatory reduction described in Section III, this means that the first, conceptual step *constrains* the second, empirical step. This point is sometimes overlooked, perhaps because empirical investigation usually precedes any explicit grasp of the conceptual constraints within which it operates. For instance, it's unlikely that early chemists thought of their task in this way: "our concept of water ties its identity to the microstructure of the actual watery stuff around here; so, to find an explanatory reduction of water, we should determine the microstructure of that stuff." Still, implicit features of the concept [water] determine which identity statements are justified, in the light of particular empirical facts. This role of the concept is revealed in arguments to show that "water = H<sub>2</sub>O" is true and necessary: we consult our conceptual intuitions to see that, if the watery stuff on Twin Earth is XYZ, it is not *water*.

Our concept of water, and the fact that the actual watery stuff around here has microstructure H<sub>2</sub>O, together entail that water = H<sub>2</sub>O. Crucially, *this entailment is conceptual*: for it is a conceptual truth about water that compositional facts (or those which scientists consider crucial, or whatever) determine its essence. Anyone who held that the actual watery stuff had microstructure H<sub>2</sub>O, accepted that this was taken by scientific authorities to show that water = H<sub>2</sub>O, and yet was unwaveringly disposed to deny that water = H<sub>2</sub>O, would not be employing *our* concept [water].

This point can be expressed using the now-familiar distinction between conceptual and metaphysical possibilities. While "water = H<sub>2</sub>O" is a necessary *a posteriori*, "metaphysical" truth, it is not a conceptual truth, since "water = XYZ" is conceivable. The epistemic difference between conceptual and metaphysical possibility derives from the fact that metaphysical possibilities depend on actual world facts, while conceptual possibilities do not. But metaphysical possibility constrains every possible world. Given that the term "water" rigidly designates the actual watery stuff, and the actual watery stuff has microstructure H<sub>2</sub>O, there is no possible world in which *water* isn't H<sub>2</sub>O. (There are, of course, possible worlds in which watery stuff *called* "water" isn't H<sub>2</sub>O.)

Conceptual truths alone license the inference from empirical facts, such as “the watery stuff around here has microstructure  $H_2O$ ”, to metaphysical truths such as “water =  $H_2O$ ”. Since metaphysical truths depend on empirical facts, they are not knowable *a priori*. Yet the step from empirical facts to metaphysical truths is itself an *a priori* step.<sup>18</sup>

This explains why identity claims require conceptually-based justification, despite the contribution of empirical facts. The extent and nature of a concept’s deferential component is a purely conceptual matter. The conceptual entailment from empirical facts to metaphysical truths highlights this fact: the truth of identity statements depends on empirical facts only because, and to the extent that, the relevant concepts sanction this dependence.

So conceptual analysis grounds the justification for explanatory reductions in two ways. First, only conceptual analysis can provide evidence regarding the respective extents of the descriptive and non-descriptive components of a concept. Second, conceptual analysis informs us as to the nature of these components. While it has traditionally been used to reveal the descriptive aspects of concepts, relevance considerations imply that conceptual facts also underwrite the non-descriptive aspects (in Levine’s terms, the referent’s non-descriptive modes of presentation). Conceptual analysis is thus needed to tell us what type of source—human experts, compositional facts, causal history, nomic relations, etc.—qualifies as authoritative about the kind at issue. The salient order here is not chronological but epistemic: empirical investigation need not await conceptual analyses, but explanatory reductions are *justified* only by virtue of these conceptual facts. So they are justified only if the results of (actual or potential) conceptual analyses (would) deem them so.

## VII. Conceivability Arguments about the Mind

We now turn to the sort of reductions which are the chief concern of many parties to this debate: mental-physical reductions. The most influential objections to reductive materialism about the mind rely on arguments from conceivability. According to one such argument, it is conceivable that one knows all of the physical facts about a phenomenal state without knowing all of the phenomenal facts about it (Jackson 1982); according to another, it is conceivable that there is a physical duplicate of me which lacks phenomenal states (a “zombie”) (Chalmers 1996). These arguments are intended to show that phenomenal properties are distinct from physical properties. Most materialists believe that such scenarios are conceivable, but deny that their conceivability presents an obstacle for reductionism. On that view, we can know that phenomenal properties are reducible to physical properties, despite the fact that our concepts of phenomenal properties differ profoundly from our concepts of physical properties. The preceding arguments show that materialists cannot simply dismiss conceptually-based arguments. In this final section, I briefly assess the effect of the CBJ thesis on the dialectic between materialists and anti-materialists.

Consider the statement “pain = physical state *c*”, where *c* may be a functional (second-order physical) state. One argument to show that this statement violates our concept of pain is the zombie argument: I have pain states; it is conceivable that there is a creature, physically identical to me, which is devoid of pain states; hence, pain states are not identical to physical states. This argument uses the conceivability of “physical state *c* is not painful” to reject “pain = physical state *c*”.

Materialists standardly respond to such arguments by noting that the parallel argument, regarding water, is invalid: while “the actual watery stuff around here does not have microstructure H<sub>2</sub>O” is conceivable, it is nonetheless true that water = H<sub>2</sub>O.<sup>19</sup> Identities are metaphysically necessary, but they need not be conceptually necessary. So, the materialist concludes, this moral applies to “pain = physical state *c*”: this is not a conceptual truth, but it may nonetheless be a metaphysically necessary truth.

I have argued that empirical facts qualify as evidence for a reduction of a kind if and only if our concept of the kind entails that they qualify as such evidence; therefore, empirical facts *conceptually* entail metaphysical facts. We can conceive that water is not H<sub>2</sub>O. But we cannot conceive that the following conjunction is true. (1) There are certain empirical facts—roughly, all of the actual watery stuff around here has microstructure H<sub>2</sub>O, and scientists individuate substances by their microstructural properties; *and* (2) there is water which lacks microstructure H<sub>2</sub>O. It is the inconceivability of this conjunction, revealed in Kripke-style thought experiments, which warrants the claim that water = H<sub>2</sub>O.

To show that phenomenal state-types are irreducible to physical state-types, then, it is not enough to observe that zombies are conceivable. As in the water case, pain may be reducible to physical state *c* even if zombies are conceivable on their own, apart from any empirical facts. The zombie argument instead rests on the conceivability of the following *conjunction*. (3) There are certain empirical facts—roughly, every actual instance of pain is perfectly correlated with a token of physical state *c*, and vice versa; actual tokens of *c* are caused by physical states which are perfectly correlated with the causes of pain; actual tokens of *c* cause physical states which are perfectly correlated with the effects of pain; *and* (4) a physical replica of me, which of course tokens my *c*-states, lacks my pain states.

To consider whether the conjunction (3)&(4) is conceivable is, of course, to run a thought experiment. According to the arguments of Section II, the results of this experiment reveal the extent and nature of our deference (to causal facts, compositional facts, or other authorities) regarding the kind at issue. And the CBJ thesis, successfully defended against the above objections, entails that these conceptual facts constrain explanatory reductions. In other words, the conceivability of (3)&(4) seems to refute the claim that pain = physical state *c*.

How might the materialist reductionist counter the argument that the conceivability of (3)&(4) shows that pain ≠ physical state *c*? The controversy over

these conceivability arguments suggests that this conjunction at least *appears* conceivable, in a sense of “conceivability” which carries with it the appearance of possibility. This raises the question of how conceivability is related to possibility. My understanding of conceivability is deeply indebted to Yablo (1993). Yablo says that a proposition  $p$  is conceivable for me just in case “I can imagine a world that I take to verify  $p$ ”. (Yablo 1993, 29) Yablo notes that this construal of conceivability allows “I do not exist” to be conceivable, since it doesn’t tie conceivability to believability, or to actual-world justification. Further, conceivability in this sense *involves*, and so need not presuppose, the appearance of possibility. These consequences support Yablo’s claim that conceivability, in this sense, provides some evidence for possibility. This is a weak claim, and one that is generally accepted, for it allows that the evidence which conceivability provides is *defeasible*. (Compare Levine 1998, 450–1).

Since conceivability provides only defeasible evidence for possibility, the materialist could argue that (3)&(4) is conceivable but not possible. Such an argument would proceed as follows.

What we can conceive is contingent upon psychological facts. Our current ability to conceive (3)&(4) is due to the limitations of our current understanding of the brain. If we understood the brain as well as we understand H<sub>2</sub>O, (3)&(4) would be as inconceivable for us as (1)&(2) is. We don’t yet possess the concepts necessary for such an understanding; perhaps we aren’t capable of possessing them. But possessing such concepts is required if the conceivability of (3)&(4) is to serve as evidence for its possibility.

On this view, the current conceivability of (3)&(4) derives from our empirical ignorance; specifically, it derives from our failure to possess the concepts required to adequately understand the empirical facts included in (3). This proposal denies that the generic “ $c$ -fiber firing” and “functional role  $f$ ” are effective as stand-ins for currently unknown physical facts. So-called “mysterians” such as McGinn (1991) hold this sort of position, with an added skepticism about our ability to gain the necessary concepts. On the non-skeptical version, we could eventually gain new concepts which would render (3)&(4) inconceivable. This claim—with or without the skepticism—is defended by an insistence that materialism simply *must* be true.

It is beyond the scope of the current paper to evaluate the arguments for materialism. Notice, however, that the above strategy rests materialism on independent arguments, and concludes from these that our conceptual resources are (currently) limited. Yet these independent arguments for materialism also depend on conceptual facts—most famously, that we cannot conceive of how the physical and the non-physical could causally interact. (Simplicity considerations of the sort discussed in Section IV provide further intuitive support for materialism.) So maintaining that we lack concepts crucial to understanding the mind

threatens to undermine arguments *for* materialism as well. If limitations in our current understanding of the brain mean that we can't trust current conceivability as a guide to possibility, then the evaluation of arguments on both sides of this debate must await our acquisition of new concepts. The dispute between the materialist reductionist and the anti-reductionist would thus reach a stalemate.

The materialist can avoid a stalemate only by providing good reason to deny that the conceivability of (3)&(4) is a mark of its possibility. To avoid question-beggingly basing this claim on materialism, the materialist must engage conceivability arguments on their own turf, by incorporating the impossibility of (3)&(4) into a larger theory that, if true, explains why (3)&(4) is currently conceivable. There are a variety of strategies available here, including:

- Show that the sort of information which a completed science of the brain will include in (3) is likely to be fundamentally different from the sort of information we now possess.<sup>20</sup>
- Argue that conceivability tests cannot possibly do justice to the enormous amount of facts which (3) includes. (cf. Churchland 1984, 96–8)
- Diagnose the apparent conceivability of (3)&(4) as the actual conceivability of each conjunct on its own.<sup>21</sup>
- Argue that (3)&(4) seems conceivable only because phenomenal concepts refer via a different process than physical concepts use. E.g., they refer “without the use of any descriptive, reference-fixing intermediaries” (Tye 1999, 713), or “by simulating their referents” (Papineau 1998, 384) or with a “semantically primitive Mentalese lexeme” (Lycan 1996, 64).<sup>22</sup>

That is, the materialist should provide a diagnosis of the current conceivability of (3)&(4) combined with the impossibility of (3)&(4). The difficulty lies in overriding conceivability intuitions, regarding (3)&(4), while exploiting conceivability intuitions which favor materialism, such as the conceptual obscurity of physical-nonphysical interaction.

Reference Deference provides the anti-reductionist with a competing diagnosis of the conceivability of (3)&(4).

While (1)&(2) is inconceivable, (3)&(4) is conceivable. This difference is not an aberration and we need not make recourse to the above strategies to explain it. For a difference in deference explains it: the empirical conjunct (1) makes a different contribution to the water case than the empirical conjunct (3) makes to the pain case. Our deference to physical facts and scientific authorities, regarding water, allows (1) to serve as the basis for the metaphysically necessary truth that water = H<sub>2</sub>O. Statement (3) does not serve as the basis for a parallel metaphysically necessary truth (that is, it fails to establish that pain = physical state *c*), because we do *not* defer to physical facts or scientific authorities regarding *pain*. This lack

of deference is evident in the conceivability of (3)&(4), since knowledge of all salient physical facts is stipulatively included in (3).

Each side can then diagnose the current conceivability of (3)&(4) in a way favorable to its own position. Moreover, each diagnosis goes some way towards preserving simplicity. Materialism preserves ontological simplicity, while anti-materialism preserves conceptual simplicity by using a disparity in folk deference to explain the asymmetry of “water” and “pain” intuitions.

Is this stalemate irresolvable? A survey of the moves and counter-moves may suggest that the parties to this debate are simply at loggerheads. For my part, I don’t see that this pessimistic conclusion is warranted just yet. The stalemate is an epistemic situation, and will be resolved only by new evidence. However, my central goal is not to predict the outcome of this debate, but rather to show that it must be decided on conceptual grounds. As I see it, the stalemate derives from a difference in conceivability judgments: the materialist has a comparatively strong intuition that, e.g., the physical and the nonphysical cannot causally interact, and a comparatively weak intuition that, e.g., zombies would remain conceivable even if we had a better understanding of the brain. The anti-materialist is in the opposite situation. These intuitions are susceptible to change, given new evidence. Likely candidates for such evidence include a new understanding of the brain, or a good non-materialist account of mental causation. And the evidence that would bring about these changes in intuition may well be empirical. My point is that the contribution of even empirical evidence lies in its effect on conceivability intuitions. In this way, the dispute over materialism turns on conceptual facts.

## Conclusion

We, the concept-possession folk, are the ultimate authorities regarding what is required to satisfy our concepts. External factors possess authority only insofar as we folk defer to them. Since conceptual analysis is the only method of revealing the extent and nature of folk deference to other authorities, accurate conceptual analyses determine justificatory criteria for proposed reductions, including proposed reductions of mental properties or phenomena.<sup>23</sup>

## Notes

<sup>1</sup> See, e.g., Bealer 1987, Stich 1992, Tye 1992, Chalmers 1996, Jackson 1998, Levine 1998, and Block and Stalnaker 1999.

<sup>2</sup> The exceptions are Bealer, Chalmers and, more qualifiedly, Jackson. (See Section V.) It will be clear in what follows that my view owes a great deal to these three philosophers. The general view I present here has a strong affinity to Bealer’s (1987) view in particular, but I cannot say whether Bealer would endorse my particular account, or would agree with my responses to recent arguments against this general view.

<sup>3</sup> Throughout the paper, I use double quotes to mention words and brackets to mention concepts.

<sup>4</sup> Note that premise 3 uses “analysis” in a factive sense; the accuracy of such an analysis is thus guaranteed. However, premise 3 becomes non-trivial when it is assumed, as it is here, that conceptual analysis is an *a priori* matter.

<sup>5</sup> While I don’t expect it to win any converts, my argument does provide some support for individualism. For it shows how some of the phenomena which motivate anti-individualism, such as the “division of linguistic labor”, may be accommodated within an individualist framework. In part because of this individualism about concept possession, I shall assume that concepts may outrun linguistic competence: an individual who lacks the appropriate dispositions, regarding the term “cat”, is mistaken about that term’s meaning, but she may nonetheless possess *a* concept which she falsely takes to be the concept expressed by “cat” in her linguistic community.

<sup>6</sup> The basis for conceptual truths provided by the Argument from Relevance is broadly similar to the basis which Boghossian (1997) develops in detail. Boghossian argues that abandoning analytic truths leads to full-fledged semantic indeterminacy: if there are no statements true in virtue of meaning (or, rather, in virtue of meaning and logical truths), then there are no determinate links between a word and its referent, and hence no determinate meanings whatsoever. My claim is this: if there are no conceptual truths, then there are no determinate links between a target property and a purported reduction of it, and hence purported reductions don’t truly reduce their targets.

<sup>7</sup> At times, Kripke seems to acknowledge that appropriate ascriptive factors must be in place in order for reference to occur, e.g.: “When the name is ‘passed from link to link,’ the receiver of the name must, I think, *intend* when he learns it to use it with the same reference as the man from whom he heard it.” (Kripke 1972, 96; my emphasis) And at least some of his apparent claims to the contrary are actually consistent with my construal of the argument. “On our view, it is not how the speaker thinks he got the reference, but the actual chain of communication, which is relevant.” (*ibid.*, 93) This statement is consistent with my reading, given that “how the speaker thinks he got the reference” may be influenced by his theory of reference, and his theory may be at odds with his dispositions to respond to particular scenarios.

<sup>8</sup> I accept a broadly Fregean view of concept individuation, according to which individuation is based on ascriptive elements and so each kind concept has unique ascriptive elements in its mode of presentation. But this is an easy requirement to meet: even Oscar’s concept of arthritis has a unique ascriptive element, viz., that it is that condition which relevant experts refer to as “arthritis”. (Which is not to say that every concept is individuated by an associated term.)

<sup>9</sup> Bealer characterizes his argument as “transcendental”: he takes the falsity of global scientific essentialism (roughly, the view that empirical facts partially determine the referent of *every* term) as a requirement for the success of local scientific essentialism (roughly, the view that empirical facts partially determine the referent of *some* terms).

<sup>10</sup> Interestingly, this acquiescence probably qualifies as free on some compatibilist “deep self” accounts of freedom. For, upon reflection, we are likely to endorse this acquiescence, since it allows for a breadth of knowledge unavailable to one who insists on performing all of the conceptual labor oneself. And “deep self” accounts allow that non-deliberate acquiescence which we would reflectively endorse can be free.

<sup>11</sup> The individualistic consequences of this view are worth emphasizing. Concepts and conceivability are linked: the boundaries of a concept are logically tied to conceivability intuitions. Only individuals can *have* concepts and conceivability intuitions. Communities can have practices, but these practices determine, at most, the actual-world extensions of a term such as “water”. The modal and intensional features of “water”, including the fact that water is identical to H<sub>2</sub>O, depend on individual dispositions to characterize hypothetical cases, and thus depend on individual concepts. This means that shared concepts are defined by, and derivative from, the individual concepts which are logically linked to such dispositions. As I said in note 5, my argument offers some support for individualism by showing how individualism can accommodate some key intuitions behind anti-individualism, such as the intuition that linguistic labor is divided.

<sup>12</sup> The two-dimensional account is not original with Jackson or Chalmers, and some of its proponents are not concerned with maintaining *a priori* philosophical methodology. Chalmers acknowledges his debts as follows. “This [two-dimensional] framework is a synthesis of ideas suggested by Kripke, Putnam, Kaplan, Stalnaker, Lewis, Evans, Davies and Humberstone, and others who have addressed these two-dimensional phenomena.” (1996, 56)

<sup>13</sup> “Even if it is a microphysical fact that H<sub>2</sub>O is *a* waterish stuff around here, it is not a microphysical fact that it is *the* waterish stuff around here.” (B&S 1999, 19) Compare my (1999a), footnote 15. They do not address the strategy I consider in that paper: using the causal closure of the physical, together with the denial of overdetermination, to conclude that nothing nonphysical fills the “water” role.

<sup>14</sup> Could one defer about a conceivably nonphysical kind? I don’t see why not. The spiritualist may defer to religious or psychic authorities about the nature of spirits; the Platonist may defer to philosophers about the nature of justice. This suggests that Step 2 needn’t even be a specifically empirical premise. I gloss over this possibility in what follows.

<sup>15</sup> The opponents of B&S who favor the two-dimensional account defend it on principled epistemic grounds like those which support CBJ. Any commitment to a particular ontological thesis about the microphysical stems from this more basic epistemic commitment. For instance, Jackson makes clear that his deep methodological point is independent of a particular ontological position. In arguing for the “entry by entailment” thesis, on which he bases his conceptual approach to reduction, Jackson says: “although the argument was developed for the special case of physicalism and the psychological, the argument did not depend crucially on matters local to that special case. We could have argued in the same general way in the case of physicalism and the semantic, or in the case of Cartesian dualism and the semantic, or in the case of Berkeleyan idealism and physical objects.” (Jackson 1998, 26)

<sup>16</sup> “The role of identities is to disallow some questions [such as *what explains this?*] and allow others [such as *how can two terms denote the same thing?*]” B&S (1999, 24). Compare Papineau (1998).

<sup>17</sup> The folk may have evidence for the existence of some kinds which is epistemically so basic that nothing can override it. Introspective evidence for phenomenal kinds is the most obvious candidate here.

<sup>18</sup> Compare Bonjour’s (1997) argument for the more general conclusion that all empirical justification depends ultimately on *a priori* justification. (Bonjour 1997)

<sup>19</sup> A strict parallel to the zombie argument would exploit the conceivability of “H<sub>2</sub>O does not have the same microstructure as the watery stuff around here”. I use the conceivability of “the watery stuff around here does not have microstructure H<sub>2</sub>O”, and the conceivability of “the watery stuff around here has microstructure XYZ” because these are familiar cases. Given that H<sub>2</sub>O is distinct from XYZ, each “the watery stuff around here has microstructure XYZ” and “the watery stuff around here does not have microstructure H<sub>2</sub>O” individually entails “H<sub>2</sub>O does not have the same microstructure as the watery stuff around here”.

<sup>20</sup> Levine expresses doubts about this strategy. “The point is, we know a lot already about how the physical world is put together, and how information can be processed by physical systems. If the conceptual tools this knowledge provides aren’t enough to bridge the explanatory gap... it’s totally unclear what else could be on the horizon.” (Levine 1998, 472)

<sup>21</sup> In a similar vein, Tye suggests that the conceivability of each conjunct, on its own, explains the apparent conceivability of the conjunction which is crucial to Block’s (1978) “Chinese-body” argument against reducing intentional states. (Tye 1995, 199–200)

<sup>22</sup> More needs to be done to show that physicalism is compatible with phenomenal concepts’ special mode of referring. Dualists can argue that the fact that phenomenal concepts uniquely refer—by simulating their referents, say—actually undermines physicalism, as follows. Simulating a phenomenal state, and hence instantiating the relevant phenomenal property, is required to truly understand the phenomenal property; instantiating a physical property is never required to truly understand the physical property; therefore, phenomenal properties are significantly differ-

ent from physical properties. It is not a sufficient explanation to say that this difference lies merely in a difference in how these properties are conceived, since it persists even if we acknowledge the different modes of reference. Denying physical-phenomenal identities provides a better alternative explanation. (This is just a sketch of a possible argument; for development of a similar line, see Gertler (2001).)

<sup>23</sup> Earlier versions of Section III were presented in April 2000 at Tucson 2000: Towards a Science of Consciousness, and at the APA Central Division 2000 meeting in Chicago. I thank the audiences there and my commentator at the APA, Al Martinich. I delivered an abbreviated version of the entire paper at the University of Wisconsin–Madison in September 2000, and received numerous useful comments. I am deeply grateful to many individuals who have generously provided comments on various sections at various points, including Felicia Ackerman, Alex Byrne, Dave Chalmers, Jamie Dreier, Steve Horst, Trenton Merricks, Jim Pryor, Sam Rickless, Ted Sider, Elliott Sober, Steve Yablo, and an anonymous referee for this journal. Special thanks go to William Lycan for provocative objections and helpful suggestions.

## References

- Bealer, George. (1987) “The Philosophical Limits of Scientific Essentialism”. *Philosophical Perspectives* 2, 289–365. Atascadero, CA: Ridgeview Press.
- Berlin, Isaiah. (1969) “Two Concepts of Liberty”. *Four Essays on Liberty*. Oxford: Oxford University Press.
- Block, Ned. (1978) “Troubles with Functionalism”. In *Minnesota Studies in the Philosophy of Science*, Vol. 9. Minneapolis: University of Minnesota Press, pp. 261–325.
- Block, Ned and Robert Stalnaker. (1999) “Conceptual Analysis, Dualism, and the Explanatory Gap”. *The Philosophical Review* 108:1–46.
- Boghossian, Paul. (1997) “Analyticity”. *A Companion to the Philosophy of Language*, eds. Hale and Wright, 331–68. Oxford: Blackwell Publishers.
- Bonjour, Laurence. (1997) *In Defense of Pure Reason*. Cambridge: Cambridge University Press.
- Burge, Tyler. (1979) “Individualism and the Mental”. *Midwest Studies in Philosophy* 4:73–121.
- Chalmers, David. (1996) *The Conscious Mind*. Oxford: Oxford University Press.
- Churchland, Paul. (1984) *Matter and Consciousness*. Cambridge, MA: MIT Press.
- Fodor, Jerry and Ernest Lepore. (1992) *Holism: A Shopper’s Guide*. Cambridge, MA: Blackwell.
- Gertler, Brie. (1999a) “A Defense of the Knowledge Argument”. *Philosophical Studies* 93:317–36.
- Gertler, Brie. (1999b) Review of Frank Jackson, *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. *Ethics* 110: 202–5.
- Gertler, Brie. (2001) “The Explanatory Gap is Not an Illusion: Reply to Michael Tye 1999”. *Mind* 110:689–94.
- Jackson, Frank. (1982) “Epiphenomenal Qualia”. *The Philosophical Quarterly* 32:127–36.
- Jackson, Frank. (1998) *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.
- Kripke, Saul. (1972) *Naming and Necessity*. Cambridge, MA: Harvard University Press.
- Levine, Joseph. (1998) “Conceivability and the Metaphysics of Mind”. *Noûs* 32:449–80.
- Lycan, William. (1996) *Consciousness and Experience*. Cambridge, MA: MIT (Bradford).
- McGinn, Colin. (1991) “Can We Solve the Mind-Body Problem?” *The Problem of Consciousness*, 1–22. Cambridge, MA: MIT Press.
- Papineau, David. (1998) “Mind the Gap”. *Philosophical Perspectives* 12:373–88.
- Putnam, Hilary. (1975) “The Meaning of ‘Meaning’”. *Mind, Language, and Reality*, 215–71. Cambridge: Cambridge University Press.
- Quine, W.V. (1961) “Two Dogmas of Empiricism”, reprinted as Essay 2 in *From a Logical Point of View*. Cambridge, MA: Harvard University Press.
- Rosenthal, David. (1986) “Two Concepts of Consciousness”. *Philosophical Studies* 94:329–59.
- Smart, J.J.C. (1959) “Sensations and Brain Processes”. *The Philosophical Review* 68:141–56.

- Sober, Elliott and Peter Hylton. (2000) "Quine's Two Dogmas". *Proceedings of the Aristotelian Society* 74(Supp.):237–80.
- Stich, Stephen. (1992) "What is a Theory of Mental Representation?" *Mind* 101:243–61.
- Tye, Michael. (1992) "Naturalism and the Mental". *Mind* 101:421–41.
- Tye, Michael. (1995) *Ten Problems of Consciousness*. Cambridge, MA: MIT (Bradford).
- Tye, Michael. (1999) "Phenomenal Consciousness: The Explanatory Gap as a Cognitive Illusion". *Mind* 108:705–25.
- Yablo, Stephen. (1993) "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research* 53:1–42