Textbook Kripkeanism & the Open Texture of Concepts

draft of July 21, 1998

Stephen Yablo

yablo@mit.edu

> [one imagines producing] an exhaustive list of all the circumstances in which the term is to be used so that nothing is left to doubt.... construct[ing] a complete definition, i.e., a thought model which anticipates and settles once for all every possible question of usage...in fact, we can never eliminate the possibility of some unforeseen factor emerging, ... and thus the process of defining and refining an idea will go on without ever reaching a final stage.
>
> F. Waismann, "Verifiability"

I. Introduction

A lot of people appear to have drawn the same "good news - bad news" lesson from their reading of Saul Kripke on conceivability. The bad news is that conceivability evidence, particularly of a "conceptual" or "a priori" sort, is highly fallible. Very often one finds a statement E conceivable, when as a matter of fact, E-worlds cannot exist. So it is, for instance, with the conceivability of water in the absence of hydrogen, or of Hesperus without Phosphorus.

The good news is that (although conceivability evidence is fallible) the failures always take a certain form. A thinker who (mistakenly) conceives E as possible is correctly registering the possibility of something, and mistaking the possibility of that for the possibility of E. There are illusions of possibility, if you like, but no outright delusions or hallucinations.

The good news is important because it gives a way of living with the bad. That a statement E is conceivable may not itself be proof that E is possible; but proof is what it becomes in the absence of an E_* such that it was really E_* that was possible, and E_* whose possibility was misread as the possibility of E.

Now, what is the relation between E and E_* whereby the one's possibility is so easily misread as the possibility of the other?

The quick answer is that E_* maps out the way the proposition that E is presented in thought; it is, for short, a presentation of E. The usual sort of presentation takes proper names in E and replaces them with descriptive and/or demonstrative phrases that, as Kripke says, fixes their reference; so, "water" might be replaced by "the predominant local clear drinkable stuff."

But the essential point is that E_* delivers the propositional content of E as a function of the circumstances that obtain where E is uttered. What E actually says, assuming the actual world is w, is the same as what E_* says about w, i.e., what it says considered as a description of w. Suppose for instance that E is "water is plentiful." Then what E actually says, pretending that the actual world is a w whose watery appearances are appearances of XYZ, is what E_* = "the clear drinkable stuff is plentiful" says about w, viz. that its clear drinkable stuff is plentiful, viz. that XYZ is plentiful.

Now, it comes as no surprise that the possibility of a presentation of E should be confused with the possibility of E . A world of which E's presentation is true is a world such that, had it really obtained, E would have expressed a truth.

But an understandable confusion is a confusion nevertheless. The possibility of "water is plentiful" expressing a truth is one thing -- it's the possibility of there being lots of watery stuff -- the possible truth of what it does express is another -- it's the possibility of there being lots of H_20.

Two notions of possibility, then. Our job as philosophers is, first, to clearly distinguish the two notions, and second, to explain how they are related. The first part is easy; it is part of the folklore of the subject that

> (i) an E that could have expressed a true proposition is "conceptually possible," while an E that does express a proposition that could have been true is "metaphysically possible"

The second part is not too difficult either. By (i), E is conceptually possible iff it expresses a truth in some w-considered-as-actual. By definition of "presentation," the truth E expresses in w-considered-as-actual corresponds to a true description its presentation E_* gives of w-considered-as-counterfactual. By (i) again, for E_* to be true of a counterfactual world is for E_* to be metaphysically possible. Hence

> (ii) E is conceptually possible iff E_* is metaphysically possible.

The philosophical payoff is as follows. From (i) we see why it is so often a mistake to infer a statement's metaphysical possibility from its conceivability. Conceivability (particularly of a conceptual or a priori sort) tracks in the first instance conceptual possibility, not the metaphysical sort. It appears from (ii), though, that the inference is not a mistake when no obfuscating presentation can be found, that is, when there is nothing to play the role of E_* but E itself. In that case (ii) tells us that E is possible in the one sense if and only if it is possible in the other.

II. Textbook kripkeanism

The story just told can be called textbook kripkeanism about conceivability and possibility. How well it corresponds to any actual belief of Kripke's is hard to say, and something I take no stand on. What I do think is that textbook kripkeanism is not right. The "good news" that E's conceivability ensures its possibility whenever no obfuscating presentation suggests itself is too good to be true.

About sixty years ago, the philosopher Charles Hartshorne put a neat twist on Anselm's ontological argument for God's existence. Granted, he said, that existence is part of God's essence does not itself show that God exists; it implies only that if God were to exist in some world, then he would exist necessarily. God in other words is either necessary or impossible. But, God is not impossible, since we can easily conceive him. Hence God is necessary, and so actual.

The orthodox response is that Hartshorne is punning on "possible." All God's conceivability establishes is his conceptual possibility. The premise needed to establish his necessity, however, is that he really could have existed. Only if there is a possible world that really contains him can we say: he exists in w, so his essence is satisfied in w, so he has the property of necessary existence in w, so he exists in every possible world, this one included.

All of this is very familiar. The reason for mentioning it is that assuming textbook kripkeanism, it fails to block the argument. Let it be that God's conceivability establishes only that he is conceptually possible. Still, the gap here is not very large. A statement's conceivability suffices for its metaphysical possibility except in those cases where all we have cottoned onto is an E_*-world passing itself off as E.

The question is: can we find a presentation of E = "there is a being whose essence includes existence" such that it

is really only this presentation that is possible, not the proposition that it presents? The presentation would replace name-like expressions in E with nonrigid descriptive phrases spelling out how we identify their referents in thought.

But, and this is putting it mildly, it is hard to think what the reference-fixing descriptions could be, or what they would replace; the statement "there is a being whose essence includes existence" seems already to be about as conceptually articulate as one could want. Another way to put it is that it is hard to see what the genuine possibility is that we mistake for the possibility of an essentially existent being. Without a separate possibility "in the neighborhood" to point to as what was confusing us, it seems we have to conclude that it is E = "there is a being whose essence includes existence" that is possible. And now it follows that a being like that truly exists.

In case anyone is not alarmed by the story so far, let me stretch it out a little. Another thing that seems clearly conceivable is that there should fail to be a being whose essence includes existence; it seems conceivable, in fact, that there shouldn't be anything whatsoever. Once again, it is hard to think of a presentation of "there isn't anything" such that it really this presentation that is possible, and this presentation whose possibility is mistaken for the possibility of emptiness.

Now we have talked ourselves into a contradiction. Textbook kripkeanism has the result that (Hartshorne's) God exists in some worlds but not in others. But it is a conceptual truth about this God that he exists in every world or none. The same problem arises for other "modally extreme" entities: numbers, pure sets, transcendent universals, and so on. Given textbook kripkeanism, they are not merely recherché, they are paradoxical. Nor can the paradox be evaded by saying that numbers and sets do not exist; it flows from the very concepts involved.

III. Consciousness

If textbook kripkeanism could be seen at work only here, in connection with God and other modally extreme entities, it might not be worth making a fuss about. But it plays a role too in an increasingly popular objection to physicalism pressed by Frank Jackson and David Chalmers.

Any physicalism worthy of the name says that the world's mental aspects are necessitated by what goes on here physically. But there is at least one sort of mental phenomenon -- consciousness -- that we can conceive going missing in a world that is physically just like ours. In a word, zombie worlds are conceivable. Doesn't this run directly against the physicalist's necessitation thesis?

Not according to most people. All that follows from the conceivability of zombie worlds is that they are conceptually possible. And it would take their metaphysical possibility to bother the physicalist.

All of this is again very old news. The effect of textbook kripkeanism, however, is to call it into question. Space between conceptual and metaphysical possibility can open up only under fairly special conditions. And, it will be said, these conditions aren't met in the present case. Zombie worlds had better be conceptually impossible, then, if physicalism is to have a chance.

Now, as it happens, Jackson and Chalmers have slightly different reasons for thinking that the zombie scenario is one where the conceptual/metaphysical distinction finds no foothold.

The crucial point for Jackson is that we are considering a world stipulated to be physically just like ours. He thinks he can get the physicalist to admit that when physical premises a posteriori necessitate nonphysical conclusions, additional physical premises can be found to make the necessitation a priori. Since in the zombie scenario we are allowed complete physical information, the additional physical premises have "already been added." So physical premises conceptually necessitate consciousness if they necessitate it at all.

What makes the zombie scenario special for Chalmers is less the nature of the (physical) premises than that of the (phenomenal) conclusion. Like Kripke, he is impressed by the fact that the way the proposition that I am in pain is presented in thought is scarcely to be distinguished from the proposition itself. To put it in terms of presentations, E_* = "I am in a state that hurts" is necessarily equivalent to E = "I am in pain." And if statements are true in the same possible worlds, then there is little prospect of explaining away the apparent possibility of one as the genuine possibility of the other.

IV. Jackson against the physicalists

The textbook kripkeanism of Chalmers's strategy is plain to see. How Jackson fits in will take a little explaining. His essential claim, remember, is that if pain is necessitated a posteriori by physical premises, then an expanded set of physical premises necessitates pain a priori.

The argument for this begins from a puzzle. At first we are inclined to think of understanding as knowledge of truth conditions: for our purposes, knowledge of which worlds a sentence truly describes. If that is the correct theory, though, then understanding a necessarily true sentence E should suffice for appreciating its necessity. And it clearly does not. I can understand "where there is H_20, there is water" without realizing its true modal status.

But the reason for my oversight is no great mystery, says Jackson. It's just that I am under- or misinformed about what chemical substance is (in the present context) picked out by the reference-fixer of "water," and to that extent aware only in a potential or hypothetical sense of the truth conditions that E in fact possesses. That this does not prevent me from understanding E suggests that understanding is a matter not of knowing the conditions under which E is true, exactly, but

> knowing how the conditions under which it is true depend on context, on how things are outside the head (1994, 39).

A little more explicitly, it is knowing the meaning function E_m mapping contexts in which E might be uttered to its truth-conditions in those contexts. Since one can grasp this meaning function without knowing E's actual truth-conditions, simply through ignorance of which context actually obtains, the puzzle dissolves. One can't be expected to see E's necessity if one doesn't know its truth conditions.

Notice what this implies, however. If it is ignorance of context that enables me to miss E's truth conditions, then once this ignorance is remedied, I am out of excuses. Semantic competence in other words should enable me

> to move a priori from... statements about the distribution of H_20 combined with the right context-giving statements, to information about the distribution of water (1994, 39).

This takes Jackson close to his desired conclusion that whatever is metaphysically necessitated by the full physical

story is conceptually necessitated by it. But a detail has been left hanging.

Why should the context-giving information be physical information? Couldn't the reference-fixer for "water" mention, say, the fact that it is supposed to be something clear and tasteless? Of course it could. But remember, Jackson says, we are asking after the consequences and commitments of physicalism. And the physicalist of all people is in no position to doubt that context is ultimately to be described in physical terms It thus appears that whatever is necessitated by physics is conceptually necessitated by it. And this applies in particular to psychology:

> the physicalist is committed to there being an a priori story to tell about how the physical way things are makes true the psychological way things are. [Note,] the story may come in two parts. It may be that one part of the story says which physical way things are, P_1, makes some psychological statement true, and the other part of the story, the part that tells the context, says which different physical way things are, P_2, makes it the case that it is P_1 that makes the psychological statement true. What will be a priori accessible is that P_1 and P_2 together make the psychological statement true (1994, 40).

Obviously though there are various psychological statements that are not a priori necessitated by physical ones, such as the statement that there is conscious experience. So, they are not necessitated by physical statements at all, so physicalism is false. That completes the argument.

V. The link with textbook kripkeanism

The puzzle that Jackson uses to disprove physicalism is really just the puzzle of a posteriori or nonconceptual necessity. Why isn't all necessity the conceptual kind? It can equally well be stated in terms of the "dual" notion of conceptual possibility, where E is conceptually possible if, roughly, it is not a priori that not-E. How can E be conceptually possible without being really possible?

Textbook kripkeanism has a view about this combination of features. The one and only way for E to be conceptually possible but not "really" --metaphysically -- possible is for something else to be really possible, namely E's presentation E_*. This presentation being an a priori equivalent of E that specifies what E says as a function of worldly context, the claim is that uttered in the right context, E would have expressed a truth.

But, this is very close to what Jackson tells us. According to him, the reason we don't see that not-E is impossible is that the meaning function E_m telling us what proposition E expresses in a given worldly context occasionally yields the result that it expresses a true proposition. Thinking of the textbook kripkean's E_* as an attempted linguistic expression of Jackson's meaning function E_m, the two stories basically agree.

VI. Knowing which

So, then: Jackson's argument is an example of textbook kripkeanism. The connection here is suggestive in both directions. Having seen earlier that textbook kripkeanism overgenerates modal "truths," e.g., it yields the contingency of theism, the suspicion is that Jackson's strategy may overgenerate as well. Having not seen earlier what the textbook kripkean's mistake in fact is, it becomes tempting to look for evidence in the Jackson argument of what might be misleading textbook kripkeans more generally. Our basic question, remember, is: how can an impossibility go unnoticed except under color of a suitable presentation, or now, meaning function?

Start with the matter of why the "contextual information" needed to boot an a posteriori necessity up into a conceptual one should be physical information. Jackson says that the physicalist of all people is in no position to deny that context is physical. But there has to be more to it than that. The physicality of context is one thing, the physicality of information about context -- the information speakers need to parlay their understanding of E into knowledge of its truth conditions -- is another.

So let us ask again: why should physicalists think that the contextual information is physical? They are not deniers of nonphysical information, after all. They merely insist that it be necessitated by physical information. If the necessitation were conceptual, then no problem; information that is conceptually necessitated by physical information can be considered itself physical. But to insist that the necessitation is conceptual would seem to beg the question at issue.

Or maybe not. Suppose that a physical description P of the context necessitates a nonphysical description Q. (P and Q might be "H20 plays such and such a role" and "H20 is water.") Then the conditional "if P then Q" threatens to be the very sort of necessary truth that Jackson says he finds puzzling. Why isn't it conceptually necessary? The only possible answer is that it has necessary truth conditions in this context, non-necessary ones the next context over.

This is reintroducing a complication we had thought to be done with. Given that P and Q were brought in to pin down the context of E enough to settle its truth-conditions, it seems only fair to allow that they do not bring with them further context-sensitivities. And now the thinker has no excuses; "if P then Q" has got to be conceptually necessary, in which case the physicalist may as well concede the context-giving information Q is indeed physical.

Notice the underlying assumption: the puzzle about nonconceptual necessities is such an extremely puzzling puzzle that it's not allowed to even exist except when Jackson's preferred strategy of solution is available. Anyone who really and truly knows which worlds "if P then Q" is true at has got to realize that it is true at all worlds. I want to flag that assumption because it's going to come up again. How does the argument fare from this point on?

Understanding E = "there is pain" is knowing how its truth conditions vary with context. The physicalist is allowing that it takes only physical information to know which context one is in, nearly enough at least to be able to compute E's truth conditions. So, someone who understands “there is pain” and possesses the relevant physical information knows which worlds are E-worlds. But (and let's flag this assumption too) anyone who really knows which worlds are E-worlds thereby knows whether the E-worlds includes all worlds physically just like this one. Putting the pieces together, anyone who really understands "there is pain" is in a position to parlay purely physical information about context into the knowledge that zombie worlds are impossible.

Both stages of the argument depend on hypotheses about what "else" ought to be known by someone who knows which worlds a statement truly describes. And indeed the puzzle itself depends on such a hypothesis; knowing which worlds a necessary statement is true of is supposed to suffice for knowing that it is true of every world. Here is the general schema:

> (+) knowing which worlds are E-worlds suffices for knowing that the E- worlds are (include, etc) the F-worlds, assuming they in fact are.

As a general requirement on knowing which, this seems like asking a lot! For one thing, I may not have a very good idea of which worlds are F. Take for instance the worlds that are physically just like this one. Unless I know which

worlds these are -- and given how little I know about the physical nature of this world it seems an open question -- knowing which worlds contain pain is clearly not going to tell me whether the pain-worlds include them. Or let the F-worlds be the class of all possible worlds bar none. If I am uncertain about which worlds are really possible (and I am) then there is nothing to prevent me knowing which worlds physically just like ours contain pain while still failing to know whether all worlds fall into this category.

But the real reason (+) doesn't work is one that applies even when we know which worlds are F. The real reason is that the standards for "knowing which" are themselves so intentional and context-driven as to prevent any easy conclusions about what the knower is now in a position to appreciate.

This much seems plausible: for me to know which worlds make E true, I need a way of picking out the E-worlds in thought, and not any old way will do. But the sort of way that suffices is not a function of the set of worlds alone. It depends on (among other things) on its being the sentence E that is used to designate the set as opposed to some necessarily equivalent alternative. I know which worlds E = "there is pain" is true of by knowing that they are the worlds in which there is pain. (If more than that is required, count me among those who don't get it.) I know which worlds F = "things are physically as in our world" describes by knowing that they are the worlds in which matters are physically as in our world -- and here I might be able to reel off some specific physical requirements. Obviously though to know in these sorts of ways which worlds E and F are true of does not put me in a position to tell whether E is true in every F-world, even if in fact it is.

VII. Canonical conception

One line of response would be to equate understanding with some sort of unmediated, perhaps acquaintance-like, grasp of which worlds make your sentence true; that will be postponed for a bit, until after Chalmers. Another is to insist that understanding a sentence is matter of knowing which set of worlds it expresses in a special canonical way: a way that better responds to what worlds in their innermost nature are.

Some such adjustment might seem called for anyway, since otherwise the equation of understanding with knowledge of truth conditions flirts with triviality. No doubt understanding "France is a democracy" goes with knowing that the worlds it is true of are the ones where France is a democracy. But this sort of explication doesn't seem to take us very far. It would be better (one might think) if the verifying worlds could be identified not as whatever makes it the case that E, but, well, as the worlds they are.

Now, since the physicalist thinks that worlds are in their innermost nature physical, he will presumably insist on a physical specification. But then it can't be that the speaker "misses" the fact that any world physically like ours is a pain-world simply through failing to think of the pain-worlds in physical terms. Thinking of them in physical terms is a condition of understanding, and we are talking about a speaker who understands.

The claim is that if physicalism is true, then to understand E one must be able to decide (i) on the basis of physical information (ii) how to make the cut between E- and non-E-world in physical terms. (If physicalism is true, then understanding is "physical" understanding.) This plugs the gap in Jackson's argument, and his conclusion is now reinstated. Whatever physical premises necessitate at all, an expanded set of physical premises conceptually necessitates. Merely to understand the sentences is to appreciate their truth-relations.

Quite right, but so what? The intuition the physicalist has got to be careful not to flout is that a normal understanding of "things are physically like so" and of "there is pain" should leave open the possibility of zombie worlds. That a physical understanding of the same sentences should leave this possibility open is not intuitive at all. On the contrary: a physical understanding of "there is pain" is by definition an ability to tell whether worlds presented in physical terms do or do not contain pain. The only physicalist who should be bothered by the refurbished argument is the one (if he exists) who thinks ordinary understanding is physical understanding as defined by (i) and (ii). And that sort of physicalist deserves to be in trouble.

Everything here goes back to the assumption that the physicalist will insist on a physical specification of the verifying worlds. Why should he? Physicalism was supposed to be an ontological theory, not a theory of understanding. This distinction is trampled on when understanding is equated with canonical grasp of truth conditions. It now becomes a "consequence" of physicalism that typical speakers, to the extent that they find zombie worlds conceivable, don't really understand "there is pain"! The physicalist presumably finds this as bizarre as anyone else. Why should one's claim to understand "there is pain" depend on such an arcane and out of the way matter as the possibility of zombie worlds?

VIII. Chalmers against the physicalists

A word first about Chalmers's semantical framework. He and Jackson agree in associating with E (as employed in a particular context) a propositional content made up of the worlds which E (as used in that context) truly describes; this content is in Jackson's terms the "truth conditions" of E, in Chalmers's terms E's "secondary intension." They agree too in assigning E an additional semantical value intended to bring out how E's interpretation varies with context.

The difference is that where Jackson's "additional" value is a meaning function from contexts to propositions (sets of worlds), Chalmers's "primary intension" is just another proposition. A world gets into E's secondary intension if E is true of that world considered as counterfactual, and into E's primary intension if E is true in it considered as actual. For short,

$|\underline{E}|_1$ = the set of E-verifying worlds, the ones making E true

$|\underline{E}|_2$ = the set of E-satisfying worlds, or just E-worlds.

Both of these intensions can be seen as arrived at compositionally from the intensions of E's component terms. The reason that "water = H_20" has a necessary secondary intension and a contingent primary one is that "water" and "H_20" agree in secondary intension only. With "water = the watery stuff," it's the other way around; the primary intension is necessary, because "water" and "the watery stuff" corefer in all worlds-considered-as-actual, but the secondary intension is not, because a counterfactual stuff (Putnam's XYZ) describable as "the watery stuff" may not be describable as "water."

To calibrate the three accounts: E's primary intension $|\underline{E}|_1$ = the set of $\underline{w}$ belonging to $\underline{E}_m(\underline{w})$ = the set of worlds in which E expresses a true proposition. (Some will recognize this as Stalnaker's "diagonal proposition.") Its secondary intension $|\underline{E}|_2 = \underline{E}_m(@)$, the set of worlds falling into the proposition that E actually expresses. The connection with Kripke is that $|\underline{E}|_1$ is the set of $\underline{E}*$-worlds, while $|\underline{E}|_2$ is the set of E-worlds. All in all, then, we have

Chalmers	Jackson	Kripke
E's primary int. $\|E\|_1$	the set of w in $E_m(w)$	the set of E*-worlds
E's secondary int. $\|E\|_2$	the set of w in $E_m(@)$	the set of E-worlds.

What is special about "there is pain" for Chalmers is that its primary and secondary intensions coincide. Unlike, say, "water is H20," the worlds in which an utterance of "there is pain" expresses a truth are the worlds at which there is pain. This is because our instinctive reference-fixer for "pain" (unlike "water") identifies its referent by a necessary and sufficient feature. Pain is the thing that hurts.

Now to the argument. If someone claims to find it conceivable that E although E is not really possible, the explanation is as follows. Conceivability intuitions track conceptual possibility, which

> comes down to the possible truth of a statement when evaluated according to the primary intensions involved...The primary intensions of "water" and "H_20" differ, so it is [conceptually] possible .... that water is not H_20. "Metaphysical possibility" comes down to the possible truth of a statement when evaluated according to the secondary intensions involved...The secondary intensions of "water" and "H20" are the same, so it is metaphysically necessary that water is H20 (1996, 132).

But this sort of story is not available for "pain is distinct from c-fiber firings" or "there are such and such physical goings-on without any pain," because

> with consciousness, the primary and secondary intensions coincide....The difference between the primary and secondary intensions for the concept of water reflects the fact that there could be something that looks and feels like water in some counterfactual world that in fact is not water, but merely watery stuff. But if something feels like a conscious experience, even in some counterfactual world, it is a conscious experience (1996, 133).

IX. "**Forget the semantics"**

Suppose though that someone disagrees (as they have done with Kripke) and says that the way the referent of "pain" is presented in thought can potentially come apart from the state itself; maybe "pain" stands for a condition of the brain importantly implicated in our suffering, a state that could in principle occur without phenomenal accompaniment?

This wouldn't necessarily bother Chalmers; his basic and underlying point, which he repeats again and again, is meant to be without prejudice to the proper semantics for phenomenal terms. The point is that we surely conceive some kind of world when we seem to conceive a zombie world; and that world constitutes a counterexample to physicalist supervenience whatever we say about the semantical issue:

> .... nothing about Kripke's a posteriori necessity renders any [conceptually] possible worlds impossible. It simply tells us that some of them are misdescribed, because we are applying terms according to their primary intensions rather than the more appropriate secondary intensions...It follows that if there is a conceivable world that is physically identical to ours but which lacks certain positive features of our world, then no considerations about the designation of terms such as "consciousness" can do anything to rule out the metaphysical possibility of the world. We can simply forget the semantics of these terms, and note that the relevant possible world clearly lacks something, whether or not we call it "consciousness"...the mere possibility of such a world, no matter how it is described, is all the argument [against physicalism] needs to succeed (1996, 134).

This is textbook kripkeanism at its purest and best: even the illusion of a zombie world is a correct perception of something, and that something is all we need to put physicalistic supervenience to rest.

X. De re and de dicto

Now, let's grant Chalmers that the difference between conceptual and metaphysical possibility is all at the level of statements, not worlds: where worlds are concerned the two sorts of possibility are really just one. His reasoning then appears strong:

(1) it is conceptually possible for there to be zombies, so

(2) zombie-worlds are conceptually possible, so

(3) zombie-worlds are metaphysically possible.

But although (2), on a natural reading, follows from (1), and (3) follows from a natural reading of (2), I wonder whether the two readings agree. The version of (2) entailed by (1) is

(2') it is conceptually possible that there be zombie-worlds.

(If you can imagine zombies, then you can imagine them plus their surrounding worlds.) But what you need to get (3) is

(2") there are conceptually possible zombie-worlds.

And the de dicto possibility of zombie-worlds asserted by (2') would seem to fall well short of the de re possibility asserted by (2").

The principal charm, as I see it, of Chalmers's procedure is that he has found a way of reaping the rewards of this de re/de dicto fallacy without actually having to commit it. He maintains, remember, that

(x) conceptual possibility "comes down to the possible truth of a statement when evaluated according to the primary intensions involved" (132).

This allows him to reach (2") directly from (1):

(1) it is conceptually possible that there be zombies, so (by (x))

(1') there are worlds in the primary intension of "there are zombies," so

(1") there are worlds which if actual make "there are zombies" true, so

(since worlds like that would seem to be all you could want in the way of a conceptually possible zombie-world)

(2") there are conceptually possible zombie-worlds.

The point is that it is (x) that saves the argument from being a straightforward modal fallacy. And if we now ask, why believe (x), the reasons turn out to be essentially Jackson's; they trace back to the assumption (+) that to know which worlds E is true in is to know a lot of other things besides. Here is how I imagine the argument going.

According to Chalmers, we can "think of the primary and secondary intensions as the a priori and a posteriori

aspects of meaning, respectively" (62). What is understanding, though, if not grasping "the a priori aspect of meaning"? It follows that what a speaker understands by E is given by E's primary intension: the worlds which, considered as actual, confer truth on E. If E is conceptually possible, that's because the speaker's understanding -- her grasp of the truth-conferring set of worlds -- leaves it open that E might be true. But, and this is where (+) comes in, it would not leave this open, if E was true in no worlds whatsoever. Hence we can be assured that E's primary intension is nonempty.

But now wait. To understand "there are zombies," I have to know that it is true in a world w iff w has such and such physical features with no consciousness. I don't have to know, though, whether that condition is satisfiable. It would be just as well, in fact, if I didn't know; any knowledge that I might have on the topic should be kept under wraps in this context. (Imagine that someone wants to test my understanding of "there are zombies" by asking which worlds it is true in; the reply "no worlds" would be ridiculous even if it were correct.) Understanding is knowing what a world has to be like for "there are zombies" to be true in it, regardless of how easy or difficult it may be for worlds like that to exist.

Here is the response I expect. Just as earlier we abstracted away from controversies about primary vs. secondary intensions, let us now abstract away from the doctrine of intensions altogether. Forget about (1) in other words; we can arrive at (2) another way. All we need is the Kripkean lesson that as far as worlds are concerned, conceptual and metaphysical possibility are one and the same. To the extent that I see no conceptual obstacle to a world -- to the extent that I find it conceivable -- I have to admit it as possible in the only sense of the word that applies. That leaves the question of course of how to describe this world. Chalmers is confident, though, that under any reasonable description, it constitutes a counterexample to physicalism.

But it is no doctrine of Kripke's that I first conceive worlds, and only later stop to ask what might be true of them. What would it be to find a world conceivable "in itself," as opposed to finding it conceivable that there should be worlds of some specified type? I take it that the latter phenomenon is the only real one, and that the talk of conceivable worlds always being possible has to be understood as code for something else: the claim that if E is conceivable then something is possible only perhaps not E itself. And that is just textbook kripkeanism, the view we are trying to find reason to believe.

XI. Why textbook kripkeanism (only) seems right

At the heart of textbook kripkeanism lies thesis (x). What is the evidence for it? Nobody doubts that a primary-intension-like notion has shown itself to have some predictive value in this area. But the inference from (1) to (1') presupposes that there is no way whatever of arranging for conceptual coherence short of including a world in the primary intension. Here is my best shot at a supporting argument.

1. E is conceptually possible. (P)
2. Understanding E leaves it open that E might be true. (1)
3. Understanding is knowing how truth depends on worldly context. (P)
4. Knowing how E's truth depends on context leaves it open that E might be true.(2,3)
5. E is true in some worldly context: some possible w considered as actual. (4)
6. E is true in w, considered as actual, iff w is an $|E|_1$-world. (Def. of $|E|_1$)

7. So, $|\underline{E}|_1$ contains at least one world. (5,6)

This at least has the right shape to advance us from de dicto to de re possibility. Trouble is, everything above it granted, line 5 doesn't follow. All we get from 4 is that my way of thinking of {w | w makes E true} leaves it open that the set might have members. And that is compatible with its being the empty set in fact.

Suppose for example that E is P&-C, where P = "everything is physically like so" and C = "there is consciousness." To understand E, it's enough to understand its conjuncts, that is, to know that P is verified by the worlds that are physically like so, and that C is verified by the worlds where there is consciousness. To know in these ways the truth-conditions of P and C does not begin to tell me whether a world verifying the first can avoid verifying the second. Once again, understanding is knowing what a world has to be like to verify a statement; how easy or difficult it may be for worlds like that to exist is another matter entirely.

XII. Immaculate conception

The gap in the argument has to do with disparate ways of conceiving the same worlds. One could close it by requiring the understander to conceive the truth-conferring worlds in a single fixed way, or, alternatively, in no way at all. The first strategy has already been tried; let me not repeat it here. The second or "immaculate conception" strategy tries to relate speakers to sets of worlds directly, by which I mean not under this or that mode of presentation. Rather than knowing a condition that the E-worlds satisfy, you "know which worlds the E-worlds are" iff you know how to recognize an E-world when you encounter it.

Encounter it where? The encounter had better not be in imagination, because worlds are imagined under descriptions and it is the relativity to description that we are trying to get beyond. The idea has got to be that plopped down in w with the mission of determining E's truth value there, I would conclude that E is indeed true. Here is Chalmers:

> What would we say if the world turned out this way? What would we say if it turned out that way? For instance, if it had turned out that the liquid in lakes was H20 and the liquid in oceans XYZ, then we probably would have said that both were water...(58)

The suggestion more generally is that the primary intension of my expression E is the mapping from worlds w to the extensions I would assign to E as an actual inhabitant of w. This will have to be a me that is idealized in various respects: computing power, mobility, ability to withstand high temperatures, and so on. But the general shape of the strategy should be clear enough.

If intensions are understood like this, then the original relativity in which I know the membership of a set of worlds under one description but not another is indeed mitigated. It is replaced though by an immanent relativity in which E's extension at a world varies according to my in-world representative's point of view.

An initial reason for this is that extensions tend to be presented in indexical terms. "Water" refers to the predominant clear and drinkable liquid around here. Hence if w has different such liquids in different places, there will no simple answer to what "water" would/should be seen as referring to in w. This is why Chalmers says that it is not worlds simpliciter that go into primary intensions, but centered worlds fitted out with a marked space-time point or a designated individual and time.

No sooner do we recognize the need for a center, though, than we notice ways in which it needs to be enriched and expanded. Some referents are identified by their psychological effects (whatever causes this sensation), so room will have to be made for aspects of the speaker's psychology. The center should probably also include some indication of which direction is left, and which right, and perhaps also what the speaker is attending to at any given moment, the figure/ground relations in her visual field, and what may be occurring to her in memory. All of these can and do figure in the interpretation of the indexical phrases by which the speaker fixes the referents of her terms.

A quite different way for perspective to intrude is mentioned in a footnote attached to the passage quoted -- a footnote which reinforces the impression of an investigator hypothetically parachuting down into a world with the mission of deciding what there falls into the extensions of his words. It sometimes happens that

> whether we count an object as falling under the extension of a [word] will depend on various accidental historical factors. A stimulating paper by Wilson (1982) discusses such cases, including for instance a hypothetical case in which druids might end up classifying airplanes as "birds" if they first saw a plane flying overhead, but not if they first found one crashed in the jungle (1996, 365).

The center thus needs to take notice of the order in which various sorts of cases are presented. And this calls to mind lots of other factors capable of influencing the agent's referential inclinations in not overtly indexical ways: her hunches at any particular point about how representative the observed cases have been, her larger theoretical and practical projects, her beliefs about which sorts of classifications are going to serve these projects, how anxious she is to avoid multiplying entities, how physicalistic she is -- the whole sorry mess of presumptions and prejudices that guide us in our application of old words to new cases.

All right, but why should this be a problem? Well, the reason for going hypothetically native was to secure for ourselves an unmediated grasp of primary intensions; the primary intension of a statement found conceptually possible would then have to contain at least one world, which world could then be used (in the case of interest) as a counterexample to physicalism.

If primary intensions are made up not of worlds per se, but worlds-as-experienced-and-theorized-from-such-and-such-a-standpoint, then this rationale springs a large leak. For it could happen that whenever w as seen from one perspective (as fitted out with one center) makes it into the primary intension of E, w as seen from another perspective does not. In that case there is no determinate fact of the matter as to the emptiness or not of E's primary intension. (To say that the primary intension determinately contains w-as-seen-from-such-and-such-a-perspective gains us nothing; our interest as modal metaphysicians is in the possibility of w as such, unelaborated.)

An example might be this. Suppose that my idealized self takes up residence in a world where events that I am inclined to call pains occur on all the same occasions as events that I am inclined to describe as c-fiber-firings. Whether I decide that "pain" and "c-fiber firing" pick out one and the same type is hardly likely to be settled by my competence with the relevant terms; a lot will depend on background attitudes about ontological economy, modal intuition, the transparency of the mental, and so on. This is clear from the great identity debates of the 1950s, when it was widely assumed that mental/physical correlations would soon be found and the question was what ontological conclusions to draw.

The claim is that it is utopian to expect unaided understanding to decide philosophically loaded questions, even given a full statement of pertinent facts -- up to, but not including of course, facts about how those very questions are to be answered. A lot is going to depend on factors that are hard to see either as semantical or factual, with

the result that a world that is counted into E's primary intension on one accounting is liable to find itself counted out under another. This seriously limits the metaphysical use that can be made of our alter egos' in-world judgments. If the dualist is allowed to claim w as a world in which pain and c-fiber firings are distinct, because that is a conclusion that a well-informed inhabitant of w could reasonably draw, why shouldn't the identity theorist be allowed to claim w as a world in which they are identical, for the same reason?

The dualist could reply as follows. Look, you may be right about some possible worlds; there is no determinate answer to whether they in themselves, as opposed to they-as-judged-from-this-or-that-perspective, are to be described in a way that favors physicalism or in a way that doesn't. But there are other worlds whose anti-physicalistic import is so clear and unmistakable that all well-informed observers are going to agree. Take a zombie-world, for instance; no one could think that pain was identical to c-fiber firings there, because that world doesn't have any pain.

But to assume that zombie worlds are indeed possible just forgets the reason we handed descriptive authority to our in-world representatives. Their role was to clear the path to a nonempty primary intension, i.e.., to a zombie world. For my representative to be told outright whether w verifies E (whether others feel pain) obviously defeats the purpose, since I would be reclaiming his descriptive authority for myself. If he is not told outright, however, then a zombie-world has no better claim to membership in |there are zombies|$_1$ than does a world like ours; after all, my representative cannot tell them apart. To the extent that the "immaculate conception" strategy buys us a world, then, physicalism is unbothered. The world might be our own, consciousness and all.

XIII. Conceivability

One thing is clear: modal intuitions are fallible, and defeasible by reference to empirical data. If textbook kripkeanism isn't the way to deal with our occasional misjudgments, what is?

I suspect that textbook kripkeanism is the best we can do, if we persist in seeing modal intuition as a capacity that is at bottom conceptual in nature. Let's distinguish three progressively less implausible versions of the conceptualist thesis.

> Extreme conceptualists say that conceptual conceivability is the only sort there is. Because conceivability is a function of concepts alone, our conceiving faculty is absolutely informationally encapsulated. The role of defeaters on this view is not to educate modal intuition -- like perceptual intuition in the Muller/Lyer case, it's quite unteachable -- but to alert us that circumstances obtain in which it is not to be trusted. Learning that the water around here contains hydrogen, for instance, doesn't make other sorts of water less conceivable; it just stops us from drawing the wrong conclusions from the same old mistaken intuition. It accomplishes that by slotting into a priori biconditionals along the lines of "if the stereotypical features of water are grounded in property BLAH, then water is essentially BLAH" to enable results contrary to what our error-prone intuitions continue to suggest.

The objection to this is phenomenological. It is not that we are forced to admit that water necessarily contains hydrogen against the evidence of modal intuition. When we learn the empirical truth our intuitions change, and what we used to find conceivable we find conceivable no longer.

> Moderate conceptualists agree that empirical information has its influence by fixing the values of preordained parameters in a priori modal conditionals. The difference is that where the extremist sees the conditionals as external correctives to intuition, for the moderate they are internal to our conceiving faculty and indeed what drives it. Very roughly, we find E conceivable to the extent that we are aware of no information to suggest via apriori conditionals that E is impossible. The role of defeaters on this view is not to overrule an

> inherently error-prone faculty, but to supply a badly served faculty -- or rather the modal schemata that the faculty relies on -- with a better quality of input.

This is certainly an improvement over extremism. But there is a problem about order of explanation. According to the moderate, we are forced by a priori schemata issuing from our concept of water to find hydrogenless water inconceivable. Arguably it is the other way around. Rather than the schema determining what we find conceivable, our faith in this (or any) schema derives from the fact that when we assume its antecedent, its consequent becomes modally intuitive. The schema is better cast as a (clumsy) post facto rationalization of a preexisting readiness to let our intuitions evolve in such and such ways under the impact of new information.

> Weak conceptualists concede that the dispositions come first, the articulated modal schemata schema second. Apart from that, though, they say, the moderate is right on the money. They are right in particular to say that modal intuition evolves under the influence of something a priori and conceptually guaranteed; their mistake is just to identify this "something" as the modal schemata, when really it is the update dispositions themselves. Anyone with our concept of water is obliged to greet the news that existing water-samples have such and such a microstructure with the same intuitional shift that we did. Of course, it is quite likely beyond our discursive powers to articulate in full detail the function from possible empirical findings to intuitional shifts that a particular concept dictates; one well-known source of perplexity is how to formulate fall-back norms, e.g., the norms telling us how to react if an aspiring natural kind concept (like that of jade) fails to pan out. The fact remains, however, that there is a conceptually determined truth of the matter about what modal intuitions a given evidential diet would/should evoke in relevantly endowed thinkers.

No doubt this again is an improvement. But the link that weak conceptualism postulates between concepts and evidential dispositions is implausibly tight; not enough room is left for the phenomenon of two people sharing a concept while differing on the proper response to evidence bearing on its application. Another way to put it is that weak conceptualism skirts dangerously close to the verificationist thesis that concepts cannot float free of confirmation-conditions. Let me focus for definiteness on the notion (pushed originally by Keynes and then in a less extreme form by Carnap) of "logical probability" relations between statements -- relations that all thinkers have got to respect, on pain of irrationality, when deciding how much credence to assign a hypothesis H given evidence E.

So, how is it that weak conceptualism comes dangerously close to logical probability? That too close an association would be "dangerous" doesn't need a lot of argument; virtually no one today supposes that there is a single objectively best epistemic response to a given body of evidence, never mind one settled by logic and concepts. This is because rational thinkers, let their concepts be as similar as you like, are still going to range widely along a number of dimensions relevant to their subsequent probabilities. They will differ in their personal evidence thresholds; in the kinds of tradeoffs they favor between simplicity and strength; in the importance they assign to avoidance of error as against acceptance of truth; in their attachment to truth as against verisimilitude; in how ontologically abstemious they are; and so on and so forth without obvious limit. They will accordingly draw different conclusions from the same evidence, blamelessly but in defiance of logical probability.

So much for the "dangerousness" of logical probability. It remains to be explained how weak conceptualism comes "close" to this dangerous notion. I will argue contrapositively that anyone against logical probability should reject weak conceptualism -- that if there can be differences in conditional credence between (rational) subjects with relevantly similar concepts, then there can be differences in conditional conceivability between such subjects, and hence differences in their subsequent modal intuitions. The reason is simply that our modal intuitions are influenced by our beliefs. Learning that Twain = Clemens, or that water contains hydrogen, I cease to find the alternative conceivable. Hence if conceptually congruent thinkers form different beliefs in response to the same evidence, they are going to differ too in what they find conceivable.

XIV. Error and Defeat

According to me, not even the weakest form of conceptualism about modal intuition has any plausibility. This brings us back to our original question about how to deal with fallibility and defeasibility. If the role of defeaters is not to overrule an incurably error-prone faculty, or correct the input to a faculty that is (when not abused) error-proof, what is it?

I find no fault with the banal suggestion that I uncover my modal errors the same way I uncover intuitional errors of other kinds: by noticing how my intuitions evolve as I become better educated, while working with others to free myself of errors and oversights that may be misleading me. Here is a first stab at how the process works:

> If X finds it conceivable that E, then she is prima facie justified in believing that E is possible. That justification is defeated if someone can provide her with reason to suspect the existence of a D such that (i) D is true, (ii) if D is true, then E is impossible, and (iii) that X finds E conceivable is explained by her failing to realize (i) and/or (ii).

Hammurabi was able to conceive it as possible for Hesperus to exist without Phosphorus only because he didn't realize that the two were identical, and (maybe also) that identicals necessarily coexist. The medievals were able to conceive it as possible for dolphins to be cold-blooded only because they didn't realize that dolphins were mammals, and that mammals have got to be warm-blooded. And so on.

Now, it is tempting to suppose that Hammurabi and the medievals were even at the time aware of certain specific issues, open to independent investigation, whose unfortunate resolution would have exposed their intuition as wrong.

But it seems truer to the normal progress of modal inquiry that the conceiver is not specifically aware of her intuition's vulnerability to its eventual defeater, until the defeater comes along and does its work. Before the discovery of genes, the thought may not have been readily available that scenarios in which animal reproduction was organized along some other, non-genetic, basis were at risk of being exposed as impossible by some experiments with peas. Before it was shown how to account for locomotion, respiration, and so on in biochemical terms, the problem with a scenario in which the property of being alive is randomly distributed over physical duplicates must have been hard to appreciate as well.

None of this is to deny that the concept of an animal, or of life, must somehow prepare the way for the eventual recognition that animals necessarily propagate their kind by way of genes, or that physics guarantees aliveness. But it is striking how unaware it is nevertheless possible to be of the vulnerability of one's intuition to what emerges, in the end, as its defeater.
All we have to go on in cases like this is a generalized and undirected sense that defeat is quite possibly on the way, and corresponding feelings of unease about the doomed intuition -- feelings that are so strong in some cases as to shift one's intuitive alliances before the defeater even arrives.

XV. Zombies

Am I the only one who feels the intuition of zombies to be vulnerable in this way? I am braced for the information that is going to make zombies inconceivable, even though I have no real idea what form the information is going to take.

Of course, as with the concept of life, there has to be something in our understanding of consciousness that "prepares the ground" for the eventual discovery that anyone just like me in physical respects must also be conscious. I guess then that there is room in principle for the project of looking for features of our concepts -- of what we understand by the relevant words -- that will prevent this discovery from ever being made.

Such a project looks a lot less realistic, however, when we realize that grasp of meaning is not a normative crystal ball telling us what modal conclusions are to be drawn from every new empirical finding, however unforeseen or unforeseeable. One could stipulate, I suppose, that a fully lucid understanding of E would "anticipate" in some way the bearing of all possible observations on E's modal status, in all possible methodological climates (etc.) But that's not the kind of understanding we have, and I imagine not the kind anybody would want.

Bibliography

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

Hartshorne, C. Man's Vision of God (New York: Harper Row, 1941), excerpted in Plantinga 1965

Jackson, F. "Armchair Metaphysics," in M. Michael & J. O'Leary-Hawthorne, ed. Philosophy in Mind (Dordrecht: Kluwer, 1994)

Keynes, J. M. A Treatise on Probability (London: Macmillan, 1921)

Kripke, S. Naming and Necessity (Cambridge: Harvard, 1980)

Kyburg, H. "Degree-of-Entailment Interpretations of Probability," in his Probability and Inductive Logic (London: Macmillan, 1970)

Plantinga, A. ed., The Ontological Argument (New York: Doubleday, 1965)

Smith, E. and D. Osherson, ed., Thinking (Cambridge: MIT Press, 1995)

Stalnaker, R. "Semantics for Belief," Philosophical Topics 15 (1987), 177-190

Waismann, F. "Verifiability," in A. Flew, ed., Logic and Language (New York: Doubleday, 1965), 122-151

Wilson, M. "Predicate Meets Property," Philosophical Review 91 (1982), pp. 549-589

Yablo, S. "The Real Distinction Between Mind & Body," Canadian Journal of Philosophy, supp. vol. 16 (1990), pp. 149-201

Yablo, S. "Is Conceivability A Guide to Possibility?" Philosophy & Phenomenological Research 53 (1993), pp. 1-42