Concepts & Consciousness

Stephen Yablo

I. The Conscious Mind is a hugely likeable book. Perceptive, candid, and instructive page by page, the work as a whole sets out a large and uplifting vision with cheeringly un-Dover-Beach-ish implications for "our place in the universe." A book that you can't helping wanting to believe as much as you can't help wanting to believe this one doesn't come along every day. It is with real regret that I proceed to the story of why belief would not come.

Almost everything in The Conscious Mind turns on a single claim. The claim is that there can be zombie worlds: worlds physically like our own but devoid of consciousness. Zombie worlds provide a counterexample to psychophysical supervenience; which refutes physicalism; which sets the stage for Chalmers's "naturalistic dualism" with its contingent correlational laws; which comes scarily close to casting consciousness as an epiphenomenal by-product of its physical basis; which startles us into a measure of open-mindededness about Chalmers's "way out," in which the by-product role is abandoned for a powerful new position as the physical's intrinsic realizer; which sets up the panpsychic speculations of the final chapters.

Zombie worlds are possible because they seem possible; one can conceive a scenario in which, with no physical provocation whatever, the phenomenal lights blink out. Any argument in this style has got, of course, to come to terms with Kripke's treatment of conceivability in Naming and Necessity. But Kripke can seem to be teaching a "good news/bad news" lesson about modal intuition with favorable implications for Chalmers's case.

II. 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 be. So it is with the conceivability of water in the absence of hydrogen.

The good news is that the failures always take a certain form. A thinker who (erroneously) 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. Conceivability is not per se proof of possibility, but that is what it becomes in the absence of an E* such that it was really E* that was possible, and whose possibility was misread as the possibility of E.

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 name-like expressions in E and replace them with descriptive and/or demonstrative phrases that, as Kripke says, fixes their reference; thus, "water" might be replaced by "the predominant local clear drinkable stuff," or (for short) "the watery stuff." All that matters, though, is that E* delivers the propositional content of E as a function of the world of utterance; what E actually says, if the actual world is w, is what E* says about w.[1]

III. These ideas can be called "textbook kripkeanism" or TK. A signal contribution of The Conscious Mind is to have laid out TK in something like canonical terms. The details are of course different. Where Kripke had two statements E and E*, Chalmers has just E, but with two meanings: a primary intension

$|E|_1$ = the set of E-verifying worlds
= the set of worlds that, considered as actual, make E true,

and a secondary one

$|E|_2$ = the set of E-satisfying worlds, or simply E-worlds
= the set of worlds that, considered as counterfactual, E truly describes.

These intensions -- the second amounting to what Kripkeans would call E's propositional content, the first to the propositional content of E* -- can be seen as arrived at compositionally from the intensions of E's component terms. The reason that "water =

[1] Better, what it says about w on a referential reading.

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."

How does any of this support TK? Well, together with the two kinds of intension we have two kinds of possibility. Conceptual possibility[2] "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" (132). So-called conceivability errors occur because this is not the kind of possibility that licenses the claim that "it could have been that E," or hence the kind that interests the metaphysician. ""[M]etaphysical possibility" comes down to the possible truth of a statement when evaluated according to the secondary intensions involved...The secondary intensions of "water" and "H_20" are the same, so it is metaphysically necessary that water is H_20" (132).

Now, if "conceivability error" reflects nothing more than the mismatch just noted, then one may wonder what all the fuss was about. The world we imagine is real enough; our mistake is only to think that it is an E-satisfying world as opposed to an E-verifying one. Which means that the textbook kripkean had it right all along. Whether we are right or wrong in conceiving E as metaphysically possible, there is a genuine possibility we are picking up on, viz. E's conceptual possibility = the nonemptiness of the primary intension. Should it happen that the primary intension is, or entails, the secondary one, even this "descriptional" sort of error is ruled out; conceivability now becomes decisive evidence that it could have been that E.

IV. How well TK corresponds to any actual belief of Kripke's is something I take no stand on. What I do think is that TK is not right. The good news that conceivability ensures metaphysical possibility in the absence of an obfuscating primary "presentation" is too good to be true.

---

[2] Chalmers calls it "logical", as in "[the] distinction between "logical" and "metaphysical" possibility stemming from the Kripkean cases" (67)). This causes some confusion, since Kripke uses "logical" and "metaphysical" essentially interchangeably.

A half a century ago, the philosopher Charles Hartshorne put a neat twist on the ontological argument for God's existence.[3] That existence is part of God's essence does not itself establish theism; it shows only that if existence were possible for God, he would exist necessarily. But God is certainly not impossible, for he is "coherently conceivable." And if not impossible, he is possible, and so by our earlier reasoning necessary, and so actual.

A response that was made even at the time is that Hartshorne is punning on "possible." What God's conceivability establishes is his conceptual possibility; the premise needed to establish his necessity is that he really could have existed. Only if there is a world w that really contains him can we say: God exists in w, so his essence is satisfied there, so he exists in every world, this one included.

All of this is very familiar; the reason for mentioning it is that given TK, it fails to block the argument. A statement's conceivability suffices for its metaphysical possibility except in those special cases where all we have cottoned onto is an $|\underline{E}|_1$-world passing itself off as $|\underline{E}|_2$. And it is hard to think where in "there is a being whose essence includes existence" we are to look for the expression whose primary and secondary intensions differ, or hence what the genuine possibility is that we mistake for the possibility of an essentially existent being.

To make matters worse, 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. Where are we to look for a presentation of "there isn't anything" such that it is really this presentation that is possible, and whose possibility is mistaken for the possibility of emptiness? (Some die-hard may want to maintain that emptiness is itself presented by way of a contingent reference-fixer whose satisfaction in w by a form of non-emptiness misleads us into thinking of w as empty. Surely emptiness is not that kind of concept.) For the same reason as before, then, we should conclude that Hartshorne's God could have failed to exist.

And now we have talked ourselves into a contradiction. TK makes Hartshorne's God metaphysically contingent; yet it is part of the concept of this God to exist in every world

[3] Man's Vision of God (New York: Harper & Row, 1941).

or none. The same problem arguably arises for other modally extreme entities: numbers, pure sets, transcendent universals, and so on. To go by TK, these worse than enigmatic. They are out and out paradoxical.

V. How do we apply TK to the case where E is "things are physically like so [insert here a full physical description of reality] but devoid of consciousness"? The strategy closest to Kripke is to say that "with consciousness, the primary and secondary intensions coincide....if something feels like a conscious experience, even in some counterfactual world, it is a conscious experience" (133).

What if someone disagrees (as they have disagreed with Kripke), insisting that the way the reference of "consciousness" is fixed can potentially come apart from the state itself? Maybe "consciousness" stands for a condition of the brain that (although implicated in our experiences) 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. We surely conceive some kind of world when we find zombies conceivable; and that world constitutes a counterexample to physicalist supervenience regardless:

> .... nothing about Kripke's a posteriori necessity renders any [conceptually] possible worlds impossible. It simply tells us that some of them are misdescribed...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 [its] metaphysical possibility...the relevant possible world clearly lacks something...the mere possibility of such a world, no matter how it is described, is all the argument [against physicalism] needs to succeed (134).

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

VI. According to Chalmers, the difference between conceptual and metaphysical

possibility is all at the level of statements; where worlds are concerned, the two come to the same. If this is granted, then the following argument looks strong:

(a) it is conceptually possible that there be zombies, so

(b) zombie worlds are conceptually possible, so

(c) zombie worlds are metaphysically possible.

But wait. Although (b), on a natural reading, follows from (a), and (c) follows from a natural reading of (b), the two readings don't agree. What (a) supports is

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

(If you can conceive zombies, then you can conceive them plus their surrounding worlds.) To get (c), though, you need

(b") there are conceptually possible zombie worlds.

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

The principal charm, as I see it, of Chalmers's argument is that he has found is a way of reaping the rewards of this modal fallacy without having actually to commit it. He maintains, remember, that conceptual possibility "comes down to the possible truth of a statement when evaluated according to the primary intensions involved" (132). From this it follows that

CP If it is conceptually possible that E, then E's primary intension contains at least one world.[4]

And CP allows him to reach (b") directly from (a):

[4] "Conceptual possibility" stands in CP for the intuitive notion; the gap between intuitive conceivability and Chalmers's theoretical apparatus has to be bridged somewhere, and I am putting the bridge here.

(a) it is conceptually possible that there be zombies, so (by CP)

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

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

(b") there are conceptually possible zombie worlds.

It is CP that saves the argument from being a straightforward modal fallacy, by guaranteeing a witnessing world. CP is also a crucial prop in the argument for textbook kripkeanism; indeed if conceptual possibility is equated with robust conceptual conceivability, it has TK as an immediate consequence.

VII. Why believe CP? 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.

Start with Chalmers's idea that we can "think of the primary and secondary intensions as the a priori and a posteriori aspects of meaning, respectively" (62). What is understanding if not grasping "the a priori aspect of meaning"? A speaker's understanding of E is thus given by the (actual-world) conditions under which E is true, as encoded in the set of E-verifying worlds. Now, clearly, that E is conceptually possible implies that a speaker's understanding of it -- her grasp of the relevant set of worlds -- leaves it open that E might be true. This would not be left open, however, if E was verified by no worlds whatsoever. So we can conclude that E's primary intension is non-empty. Explicitly:

1. E is conceptually possible. (P)
2. The speaker's understanding of E leaves it open that E might be true. (1)
3. Understanding is knowing truth-conditions: how truth-value depends on worldly context. (P)
4. Knowing how E's truth-value depends on worldly context leaves it open that E might be true. (2,3)
5. E is true in some such 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, $|E|_1$ contains at least one world. (5,6)

This at least has the right shape to advance us from de dicto to de re possibility. The trouble is that, 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. Obviously though to know in these sorts of ways the truth-conditions of P and C does not even begin to tell me whether a world verifying the first can avoid verifying the second. 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.

VIII. The gap in the argument has to do with disparate ways of conceiving the same collection of worlds. One response would be to equate understanding with some sort of unmediated grasp of the verifying worlds; see below. Another is to concede that these worlds have got to be conceived under a description, but to constrain the type of description so that opacity phenomena cannot arise. Understanding E is knowing its verifying worlds in a special canonical way -- a way that respects what worlds in their innermost nature are.

Now, since the physicalist thinks that worlds are in their innermost nature physical, she will presumably insist on a physical specification. How then can it be claimed that speakers "miss the fact" that any world physically like so contains consciousness through failing to think of the consciousness-worlds in physical terms? Thinking of them in such terms is a condition of understanding, and we are talking about speakers who understand. Conclusion: if physicalism were correct, and understanding were "physical" understanding, then merely to understand P & ¬ C would be enough for the realization that it could not possibly be true.

True, but so what? The intuition the physicalist must not flout is that a normal understanding of P and C leaves open the possibility of zombie worlds. That a physical understanding of these sentences should rule out zombie worlds is not counterintuitive in the least; for a physical understanding of C is by definition an ability to tell whether worlds presented in physical terms contain consciousness.

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 speakers (even anti-physicalistic ones!) don't know the meaning of their own word. Why should anyone's claim to semantic competence hang on the outcome of an arcane modal debate?

IX. The second or "immaculate conception" strategy tries to relate speakers to sets of worlds directly, by which I mean: not as the worlds meeting such and such a condition. Rather than knowing a condition on the E-verifying worlds, I must know how to recognize such a world when I encounter it.

Encounter it where, though? Not in imagination, for worlds are imagined under descriptions, hence not immaculately. 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; the primary intension of an expression F is the function taking w to the extension I would assign F as an inhabitant of w. ("If it had turned out that the liquid in lakes was H_20 and the liquid in oceans XYZ, then we probably would have said that both were water" (58).) This will have to be a me that is idealized in various respects: computing power, mobility, physical strength, and so on. But the shape of the strategy is clear enough.

For the strategy to work, my in-world representative's descriptive inclinations need to be a function of his (= my) concepts, and not extraneous "nonsemantical" factors.[5] An example taken from Mark Wilson suggests this condition may not be met: what we count

> as falling under the extension of a [word may] depend on various accidental historical factors...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 (365).

This calls to mind lots of other considerations capable of influencing a speaker's referential behavior: her hunches about how representative the observed cases have been, her larger theoretical and practical 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

[5] Two other worries I am skipping over. Do I have a priori cognitive access to ideal-me's reactions? And doesn't "access" of any sort entail a departure from immaculateness?

our application of old words to new cases.

Suppose my idealized self takes up residence in a world where events he is inclined to call pains co-occur with events he would describe as c-fiber-firings. Do "pain" and "c-fiber firing" pick out the same type for him? Unaided understanding cannot decide such 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. (Remember the great identity debates of the 1950s, when it was assumed that mental/physical correlations would soon be found and the question was what ontological conclusions to draw.) This seriously limits the dialectical use that can be made of our alter egos' in-world judgments. If the dualist is allowed to claim w as a world where pain isn't c-fiber firings, because that is a conclusion that speakers could reasonably draw, why shouldn't the physicalist be allowed to claim it as a world where they're identical, for the same reason?

The anti-physicalist could reply that 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 (my alter ego aside, let's say) doesn't have any pain.

But to assume that zombie worlds are indeed possible forgets the reason we handed descriptive authority to our in-world representatives in the first place. Their role was to clear the way 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) just returns authority to myself, which obviously defeats the purpose. If he is not told outright, though, then a zombie-world has no better claim to membership in $|\text{there are zombies}|_1$ than does a world like ours; my representative cannot tell them apart. To the extent that the strategy buys us a world, then, physicalism is unbothered. The world might be our own, consciousness and all.