The Conceivability of Naturalism

Crispin Wright

A central dilemma in contemporary metaphysics is to find a place for certain anthropocentric subject-matters—for instance, the semantic, moral, and psychological—in a world as conceived by modern naturalism: a stance which inflates the concepts and categories deployed by (finished) physical science into a metaphysics of the kind of thing the real world essentially and exhaustively is. On one horn, if we embrace this naturalism, it seems we are committed either to reductionism: that is, to a construal of the reference of, for example, semantic, moral, and psychological vocabulary as somehow being within the physical domain—or to disputing that the discourses in question involve reference to what is real at all. On the other horn, if we reject this naturalism, then we accept that there is more to the world than can be embraced within a physicalist ontology—and so take on a commitment, it can seem, to a kind of eerie supernaturalism. John McDowell (1994) has proposed a distinctive, intendedy

Ancestors of this material were presented at colloquia in St. Andrews and Glasgow in 1996, at the Logic and Language conference held at the University of London in April 1998, at the Language and Mind seminar at New York University in that same month, when Ned Block was commentator, and at two seminars given at Ohio State University in April 1999. My thanks to the discussants on those occasions for many helpful criticisms and questions, and in particular to Paul Boghossian, Jim Edwards, Bob Hale, Robert Kraut, Neil Tennant, and Ned Block. I am also greatly indebted to the critical feedback of the editors of the present volume, Tamar Szabó Gendler and John Hawthorne.
‘non-eerie’ accommodation, involving our habituation into a more ‘relaxed’ conception of what should rank as natural, a conception of Nature which would be hospitable to meanings, to ethical and other norms, and to psychological properties. But the position he proposes—I am speaking just of my own reaction—can too easily seem more like a triumphant reaffirmation of the common-sense categories at issue than a real response to the metaphysical dilemma they pose.

This problem provides the background to the present chapter, rather than its topic. Its topic is the famous argument, outlined in the third lecture of Saul Kripke’s Naming and Necessity, that pain—and sensations generally—cannot be anything physical. What gives this argument its interest in the context of the concerns of the present volume is the manner in which it draws on considerations of (apparent) conceivable to substantiate a metaphysical conclusion. In the first part of what follows, I shall outline the background to the argument, develop its detail somewhat, and sustain it against what are, according to my understanding, the two most influential received objections to it. Then I shall make a case that, if good at all, it should generalize to cover not just sensations, but all items falling within the extensions of (in a sense to be explained) transparent concepts, with colour concepts (an example we shall stalk throughout) and secondary-quality concepts generally the obvious next port of call. At that point, the dialectical situation will be that, to the extent that the distinctive concepts of semantic, moral, and psychological discourse also approximate the relevant model of transparency, the Kripkean argument presents an outstanding challenge to any form of reductionist reaction to the metaphysical dilemma. But I do not think that dialectical situation is stable. My concluding suggestion will be that there is still an outstanding objection to the overall strategy of the argument: the assumption that drives it, that counter-conceivability is a defeater of claims of metaphysical necessity of all kinds, both a priori and a posteriori, stands in need of a (to the best of my knowledge) unremarked form of qualification. If this is right, the argument is balked even for the basic case of sensations, and the fascinating prospect of a wide-reaching exclusion of physicalism on purely conceptual grounds evaporates.

1 Kripke’s Argument

1.1 Natural Kinds and Natural Kind Concepts

Are colours natural kinds? The philosophical question, of course, is not—in the first instance, anyway—about the actual constitution of coloured things,

---

1 One philosopher who has argued so, in a much travelled talk some years ago, is again Kripke. An account of his arguments may be found in Johnston (1992).
but about the *concept* of colour. According to the usual template, a concept is a natural kind concept if, roughly, its extension is standardly explained by reference to indicators whose status as such is viewed as contingent, and if we conceive of the real determinant of the extension as a natural property, presumed to be explanatorily associated with the indicators, of whose character we may have—and anyway need—no clear idea in ordinary commerce with the concept. The concept of water, for instance, is characteristically explained by reference to the indicator properties: tasteless, colourless liquid, occurring naturally in lakes and rivers, satisfying thirst, essential to life, solvent for many substances, and so on and so forth. But if, as is usually supposed, *water* is a natural kind concept, then these marks serve not to define it but merely as pointers to an underlying natural essence—to the best of our present knowledge, that of having the chemical constitution H₂O—whose instantiation is what canonically determines whether or not a sample is water. If *water* is a natural kind concept, the indicators serve merely as *reference-fixers*: the concept of water may be glossed as, roughly, that kind of stuff whose being the kind of stuff it is explains its characteristic satisfaction of the indicators in question. A natural kind concept thus incorporates an assumption: that there is an underlying natural essence which discharges—near enough, often enough—that explanatory role. This assumption may be wrong. In that case, the concept will suffer from reference failure. Colours themselves are natural kinds if colour concepts are natural kind concepts and they are not so afflicted.

All this is familiar. The idea that many of our general, pre-scientific concepts are concepts of this kind came into prominence with Putnam and Kripke, and contrasts with an older model, associated (on no clear evidence, actually) with the later Wittgenstein, according to which the indicator properties do not bear a contingent relation to an underlying determinant of the extension of the concept, but determine that extension intrinsically, after the fashion of a cluster of *criteria*—in the specialized sense of the term that arose in the first generation of commentary on the *Philosophical Investigations*. Clearly there could be concepts—let’s call them * criterially governed* concepts—for which this model was correct. Even if our actual concept of water is indeed a natural kind concept, we might have employed instead a concept—*schwater*—for which the water indicators did play a criterial rather than merely a reference-fixing role. To be *schwater* would just be to satisfy (enough of) the (more important) indicator properties.

---

2 Except where no ambiguity arises, I shall use boldface type to indicate reference to a concept.
3 I am going past issues to do with whether such a concept could *survive* reference failure.
4 The *loci classici* are Putnam (1975: chs. 8 and 10–12), and Kripke (1980: lecture III).
How is it manifest which is our actual concept? Well, if 'water' expressed a natural kind concept, then should it turn out that there is no interesting, explanatorily unifying property underlying the presence of the water indicators in enough cases, we ought to regard the case as one of reference failure: water, like phlogiston, would be a fiction, exploded by science. But if the concept expressed by 'water' were criterially governed—if it were schwater—no such conclusion would be warranted; the use of the concept would be infeasible by any such empirical scientific development. That's only one type of case, though. What about the other, when a criterially governed concept is such that science does nevertheless disclose a natural property underlying the characteristic co-manifestations of the relevant criteria, so that a corresponding natural kind concept, had we employed it, would have been successful? That would presumably be the actual situation if the concept expressed by 'water' were schwater: schwater would turn out to be H₂O. What in that case would show that the concept was nevertheless criterially governed?

It might be thought that the difference would emerge in our attitude to certain counterfactual conditionals, for instance:

Had it turned out that there was no underlying explanatorily unifying property, we would have regarded 'water' as failing of reference.

We will affirm that counterfactual, the natural thought would be, if, but only if, our concept of water is a natural kind concept. But this isn't good enough. For the counterfactual could be wrong even if the concept expressed by 'water' is a natural kind concept—provided that, were nature to have let us down in this way, we would then have changed the concept expressed by 'water' in the light of that discovery—falling back on the criterially governed analogue, as it were. And of course the question, whether, in continuing to regard the term as referential, we would have assigned to it a concept of a different status, can be answered only in the light of some independent determination of how possession of one or the other status would show. So the counterfactual account is no help—it needs back-up by the very thing it purported to provide.

Kripke's discussion offered a different answer, in terms of modal intuition. If water is a natural kind—say, H₂O—then it is essentially that kind: something which manifested all the indicators but was not so constituted would not be water but some other kind of stuff. By contrast, if water is actually criterially governed, then such a substance would fall under that concept, so would be water, whatever its constitution, and we ought to allow that, while water is normally made up of H₂O, it could be composed of something else. So if water is a natural kind concept, and it is true that water is H₂O, it is necessarily true that water is H₂O. But if water is a criterially governed concept, it is
contingent what constitution water has—or indeed, whether it has any uniform or typical constitution at all. To determine the status of our concept of water, then, we may check our intuitions about claims such as:

Water is H₂O, but it might not have been.

Or:

Water might have had no typical physical constitution.

If water is indeed a natural kind concept, these should impress as conceptual solecisms. For natural kind concepts distinctively sustain certain forms of necessary (a posteriori) claim of which these are violations.

In what follows, I am going to assume that this proposal is broadly adequate: that the difference between natural kind concepts and others can be found in the kind of modal distinctions which Kripke highlighted.

1.2 Primary and Derivative Natural Kind Concepts

A further qualification is wanted, however. Intuitions about the necessity of identity statements have to draw on assumptions about the status of both the configured terms—for instance, in the particular example, it is taken for granted that ‘H₂O’ is itself an expression rigidly denoting a natural kind. In general, it is obvious that the class of natural kind concepts that fit the Putnam–Kripke template—concepts that purportedly denote an essential underlying property targeted by surface reference-fixers—is a secondary class, adverting by its very characterization to a contrasting background class of elite concepts—usually assumed,³ tacitly or otherwise, to be those of developed physical science—by means of which such underlying essences may be identified canonically. For modern naturalists, indeed, the natural coincides with the physical as (best) physics understands it. Call these elite background concepts, whether or not exclusively physical, the primary natural kind concepts: concepts of which it is independently given that they demarcate—if anything—what are properly regarded as natural divisions and substances, properties and stuffs.⁶ A derivative natural kind concept will then show itself by sustaining⁷ an a posteriori, but necessarily true, identity statement whose other term expresses a primary natural kind concept.

---

³ Though not, I think, by Kripke.
⁴ Naturally, the identification of these concepts will be hostage to the fate of contemporary scientific theory. We take it that ‘H₂O’ expresses a compound such concept; but the notion that hydrogen and oxygen are primary kinds is of course empirically defeasible.
⁵—or at least: requiring that there be, if it is instantiated—
A proponent of the thesis that colour concepts are natural kind concepts thus in principle has the option of maintaining that they are primary natural kind concepts. I take that to be the view of the so-called Simple Theory of colour.\textsuperscript{8} But I shall not return to that idea in the present discussion. The more usual way of taking the thesis is that, like \textit{water}, colour concepts are derivative natural kind concepts: that our understanding of the concept \textit{red}, for instance, tacitly calls for an identification:

Redness is the property thus-and-such,
whose modal and epistemic status will be exactly comparable to that of:

Water is $\text{H}_2\text{O}$.

1.3 \textit{What is the Importance of Natural Kind Concepts?}

If many of the concepts for which the criterial model might initially seem attractive are in fact natural kind concepts, that is something which it is as well to know. And the adjustment will call in turn for some re-configuration of our ideas about what is involved in thinking thoughts in which such concepts are constituents—a re-configuration closely analogous to that involved in dropping the description theory of proper names, even in its most sophisticated forms, and accepting that such expressions typically facilitate the thinking of thoughts \textit{de re}, thoughts directly targeted upon specific objects, rather than mediated via complex descriptive conditions. In short, issues are at stake here—as Kripke saw—about the character of the relation between a thought and the objects and properties it concerns.

But those issues are not our present business. Our agenda is set, rather, by a line of thought that begins with a certain simple picture of the nature and limits of natural science. The picture has it that natural science just is the empirical theory of natural kinds and their functional relationships; the project of natural science is to taxonomize what kinds of thing naturally occur and to describe how they are causally and explanatorily related, to detail the laws to which they are subject. If this is right, then—here is the line of thought—the spectre is raised that a concept that is not a natural kind concept—but that is nevertheless, intuitively, a concept of a \textit{kind} of thing, state, or event—may go on to prove, in a certain sense, \textit{scientifically recalcitrant}, and any truths that it is needed to express may consequently lie outside the domain that natural science can illuminate. In short, at least if the simple picture is accepted, the demarcation of important groups of kind concepts that are demonstrably not natural kind

\textsuperscript{8} See Campbell (1993).
concepts may be just what is needed, for those philosophers inclined to want to do so, to challenge the physical naturalism that for so many has come to seem like common sense.

Any such challenge, however, will need to surmount the simple distinction touched on above. That a criterially governed concept should be intractable for physical science is certainly possible in one kind of case: namely, where it actually has a \textit{physically heterogeneous} extension. In that case, since there will be nothing uniform at the level at which physical science operates, in which the instantiation of the concept consists, there may be no physical laws connecting its instantiation with the instantiation of other physical properties.\footnote{This is only a possibility, of course. The heterogeneous instances of a concept may still divide up into a manageable variety of physical types, each associated in parallel law-like ways with other kinds of physical states and events.} What, though, if a criterially governed concept does have a physically unified extension, nevertheless, as \textit{schwater} actually has? In that situation we could still affirm the identity, that schwater \textit{is} \text{H}_2\text{O}, since the term ‘schwater’ has a complex descriptive content—to be cashed out in terms of the indicator properties—of such a kind as to allow the identity to be true as a matter of \textit{contingency}. The force of the identity would be that

the satisfier of (a certain descriptive condition somehow factoring in the water indicators) = \text{H}_2\text{O},

and this could be a contingent truth because of the non-rigidity of the term on the left-hand-side. So schwater—the actual stuff—would still \textit{be} a natural kind—namely, \text{H}_2\text{O}—even though the \textit{concept} of schwater was not that of a natural kind; and mention of schwater, so use of the concept, could correspondingly occur in (low-level) scientific generalizations and explanations.

This is a reminder, then, that there is—of course—no \textit{direct} way of drawing conclusions about the limitations of science from the status of particular concepts. It will be one thing—if indeed, it can be done—to make a case that colour concepts, say, are not natural kind concepts; but no immediate conclusion is to be expected about the physical-scientific role of colour. The interesting question this raises is: what sort of additional philosophical argument could encompass the stronger conclusion?

\subsection{The Counter-Conceivability Principle}

In \textit{Naming and Necessity}, Kripke deployed his new apparatus of rigid designation, \textit{a posteriori} necessities, and the rest to outline a new argument—though
he did not himself endorse it—against all possibility of the physical reduction of sensation, focusing on the case of pain.

Recall that any statement identifying the essence of a natural kind will be, if true at all, necessarily so. If water is a natural kind concept, and it is true that water is $H_2O$, then it is necessarily true. It follows that evidence against the necessity of such an identification is evidence against its truth. But what should count as evidence against its necessity in the first place? Kripke’s discussion turns on a major assumption on which we should pause: that all purportedly metaphysically necessary statements, even those—of constitution, identity, or origin, for instance—whose justification is a posteriori, are hostage to what we can, to borrow Descartes’s happy phrase, clearly and distinctly conceive, for—short of its actuality—a clear and distinct conception of a situation is the best possible evidence of its possibility. This principle—the Counter-Conceivability Principle—invites us, of course, to provide an account of when a conception should rank as relevantly clear and distinct. But without taking that issue on, we can cash the principle’s operational content as being that, if one has what at least appears to be a lucid conception of how it might be that not-$P$, then that should count as a good, albeit defeasible, ground for its not being necessary that $P$. By the Counter-Conceivability Principle, all putative metaphysical necessities, even a posteriori ones, thus have to face the tribunal of what we can, as we think, clearly and distinctly conceive; and their defeat may consequently be a priori, even if their sole possible form of justification is not.

Is the Counter-Conceivability Principle correct? Naturally, it is uncontroversial for conceptual necessities: any claim that a certain truth is conceptually necessary has to be answerable to what we can coherently and lucidly conceive. But to suppose that all absolute necessities, a posteriori ones included, are subject to constraints of conceivability may seem to be at best a substantial

---

80 The Counter-Conceivability Principle is something that might really have deserved the title of ‘Hume’s Principle’, now of course purloined by the neo-Fregean programme for the foundations of arithmetic. Recall *Treatise*, I. ii. 2: “whatever the mind clearly conceives, includes the idea of possible existence, or in other words, that nothing we imagine is absolutely impossible.”

81 Kripke does not explicitly articulate the Counter-Conceivability Principle, but it is striking how naturally it seems to have come to him implicitly to assume it. It did not occur to him to respond to the hypothetical objector who thinks she can conceive of Hesperus turning out to be other than Phosphorus, or of heat being something other than molecular motion, by saying: ‘So what? What has conceivability to do with it? I didn’t claim these things were conceptually necessary!’ Rather, the validity of the objector’s prima-facie conceptions, and their prima-facie relevance, are straightaway conceded. The defence is rather that they are not what they appear to be—that what is actually conceived in these cases goes no further than *qualitative* similarity to what was intended. But I anticipate.
epistemological claim, in need of a correspondingly substantial defence. Indeed, at worst it would appear the merest blunder: the retention of an intuitively conceptually epistemology of modality for a range of cases where modal status originates not in the character of concepts at all, but in underlying essences that may go quite unreflected in our concepts of the items whose essences they are.

That is one concern. But in addition, if accepted, the Counter-Conceivability Principle may seem to enforce an objection to the necessity of certain kinds of statement that, if they configured natural kind concepts fitting Kripke’s idea of them and were true, would have to be necessary. For instance, it may seem readily conceivable that water might have turned out to have a very different chemical constitution to the one it actually has. So it cannot be necessary that water is H₂O, even if it is true that it is. And indeed, since the point will generalize, it may look as if no such identity statement—no identity statement linking a derivative natural kind concept with a specification of the purported essence of its instances—will pass the test. So the Counter-Conceivability Principle, dubious in any case, might also seem to be inconsistent with Kripke’s own view of the modal status of such statements.

Kripke’s argument about pain is set up by a lemma constituted by his response to this objection. What is wrong with the objection is its premiss. It is no doubt conceivable that scientific investigations with the same physiognomy as those that disclosed that water is H₂O might have had a different upshot—that an investigated substance, displaying all the surface indications of water, might have turned out to be XYZ instead. But then, Kripke rejoins, it would not be water that would have turned out to be XYZ. The problem for the objector is how to characterize the stuff of the imaginary scientific investigations. The characterization can hardly proceed except in terms of satisfaction of the characteristic indicator properties of water. But how exactly? If as the actual satisfier of those indicator properties, then that is a rigid designator, which will therefore produce a necessary identity when linked with ‘H₂O’. And the suggestion that it is nevertheless prima-facie conceivable that the stuff which actually has those properties, namely H₂O, might not have been H₂O is counter-intuitive indeed. What is conceivable is that something that satisfies the indicator properties, or the most widespread satisfier of the indicator properties, might have turned out not to be H₂O—but that is perfectly consistent with the necessity of water’s being H₂O.

Essentially the same point addresses the first, more general concern. Any a posteriori necessity, N, will be associated with a seemingly intelligible imaginative scenario in which the a posteriori investigations that confirm it turn out to disconfirm it instead. So much is a consequence of those investigations being
a posteriori. But it does not follow that N will be prima-facie counter-conceivable unless it is granted that the imaginative scenario involves an appropriate play with the very concepts configured in N. And that is not granted. If water is a natural kind concept, then which concept it is depends on what is the essence of water. The impression that the Counter-Conceivability Principle is all at sea as soon as necessities originating in rebus are countenanced turns on the tacit assumption of a separation between concept and essence: that, as it was expressed above, ‘underlying essences . . . may go quite unreflected in our concepts of the items whose essences they are’. That assumption simply misunderstands what is being proposed about the character of the relevant concepts that feature in necessities of identity, origin, and constitution.

The Counter-Conceivability Principle is thus under no immediate threat. But the manner of its defence may raise a concern about its utility. Will not the principle be impotent if a presumed necessity that is apparently open to counter-conception can always be excused by charging that an objector’s scenario fails genuinely to involve the relevant concepts? Indeed, it will, unless such excuses are required to be backed up by a properly principled explanation of what the alleged shortfall consists in. If water is the concept of a physical natural kind, such an explanation will be available when the scenario fails to distinguish between instantiation of that kind, whatever it is, and presentation of the indicators. The crux of Kripke’s argument about pain is that no correspondingly principled excuse is possible: that a counter-conceivability challenge against the necessity of any particular physicalistic identification of pain actually succeeds—or at least, stands undefeated.

1.5 The Argument about Pain

Suppose C-fibres are a kind of nervous pathway actually occurring in human beings. And let the proposal be that pain is C-fibre stimulation. Still, it seems readily conceivable that physiological investigation might have found no C-fibre activity in subjects suffering pain, or even that, without change in the range of our sensory afflictions, we might have lacked C-fibres altogether. So it cannot be necessary that pain is C-fibre stimulation. Ergo, since it would be necessary if true, it cannot be true. But this would go for any purported physical identification of pain. So pain isn’t the concept of a physical natural kind.

Why is this argument any better than that about water? Because in this case the premiss really does seem to be conceivable. Suppose we try to block the argument as before. So we put the question: how is the putative scenario to justify characterising the imagined scientific experiment as one in which it turns out that no C-fibre stimulation takes place in a pain-afflicted subject? Well,
what are the relevant indicators? Just one, presumably: namely, a distinctive form of discomfort. But then the difference is that to conceive of something that satisfies the indicator is, in contrast with the case of water, already enough to conceive of pain. There is a potential difference, provided by the concept, between a substance’s giving the indications of water and its being water; but there is no more to a sensation’s being pain than its giving the indications of pain—that is, hurting. So while there is an epistemic gap between conceiving of something that satisfies the indicators of water and conceiving of water, there is no such gap in the case of our basic concepts of sensation. The claim to have counter-conceived the identity of pain with C-fibre stimulation thus stands undefeated.

1.6 A Naturalist Response

But has a physicalist any cause to object to the argument so far? After all, the conclusion, properly understood, is only that pain isn’t a physical natural kind concept. Even if that is good as far as it goes, isn’t the striking conclusion that Kripke himself proposed, that pain is no physical kind of thing, out of reach? As we noted, it is consistent with its being perfectly proper to conclude that pain is not a physical natural kind concept to reserve the thought that pain might yet be a physical natural kind. The position would be exactly analogous to the case for water, the stuff, and schwater, the criterially governed concept. The extension of such a concept may always be a natural kind—as indeed it presumably is in that case. So it might be with pain, pain and ‘pain’. If pain were a criterially governed concept, then anything that satisfied the indicators of pain—that is, felt uncomfortable in the distinctive way—that indeed count as pain, and ‘pain’ would be thereby apt to feature in true contingent identities in which it was linked with physical terms.

Why do I say that the conclusion of the argument, properly understood, is only that pain isn’t a physical natural kind concept? After all, if the identity statement

Pain is C-fibre stimulation

is, if true, necessarily true, then a standing objection to its necessity has to be reckoned a standing objection to its truth; and since ‘C-fibre stimulation’ is, in effect, just a place-holder for any physicalist reduction, that is therefore a standing objection to any such reduction. The conclusion of the argument properly concerns pain, and not just pain.

But this is to forget that the argument is driven by a more complex conditional than the objector implicitly allows. What we have is not

If pain is C-fibre stimulation, then it is necessary that it is,
but

If pain is a natural kind concept, then if pain is C-fibre stimulation, then it is necessary that it is.

So, if the major antecedent is discharged, pain’s actually being C-fibre stimulation (or whatever else) will be consistent with its being so as a matter of contingency. Sure, in that case ‘pain’ would not be a rigid designator, just as ‘water’ would not be if it expressed schwater. And Kripke’s argument against physicalism assumes that ‘pain’ is a rigid designator. But where did that premiss come from? If it is put up for reductio that pain is a physical natural kind concept, then that, to be sure, enjoins that its expression, ‘pain’, is rigid. But once the argument is allowed to proceed to the point where the first assumption is discharged, then—unless some independent argument on the point is supplied—it is open to the physicalist to fall back on the view that pain is, rather, a (very simply) criterially governed concept, that ‘pain’ is consequently flexible with respect to its reference among physical kinds, and that the identity of pain with any particular neural state is consequently a possibility. And this is just the classical—as we may call it—‘Australian Rules’—physicalism about the mental, originally proposed by writers such as Jack Smart, which the Kripkean framework has been standardly regarded as squeezing out.

1.7 Rebuttal

This line of naturalist resistance arguably fails, however, and the manner of its failure is instructive. Let’s for a moment go along with the idea that ‘water’ expresses a criterially governed concept—that is, schwater. ‘Water’ accordingly has as its sense a complex descriptive condition, fashioned out of the relevant criteria, its possibly constitutionally variable referents so qualifying by dint of their satisfaction of that condition. Nevertheless, water itself is a stuff rather than a state, so there has to be a contrast between the sense of the term, ‘water’, standing for that stuff, and that of the description, ‘the state of satisfying the criteria for being water’, standing for the state that the stuff distinctively occupies. Is the description rigid or flexible? Presumably, since the criteria for a criterially governed concept are essential to it and thus invariant, the state of satisfying them will likewise be invariant. So the latter description—‘the state of satisfying the criteria for being water’—should be rigid in any case: that is, should denote the same state in talking of any possible world, even if ‘water’, whose reference in the present scenario is to whatever is in that state, is not.
What follows is that, in order for there to be any chance of assimilating the
function of ‘pain’ to that of ‘water’ as currently conceived, so that it can serve as
a correspondingly flexible designator, there has to be a corresponding distinc-
tion between the use of ‘pain’ and that of the description, ‘the state of satisfying
the criteria for being in pain’. For the latter, by an analogue of the argument just
given, will be rigid. But here’s the point: there is no such distinction. ‘Pain’, unlike
‘water’, does denote a state; moreover, and crucially, if the (single) criterion
for being in pain is that mooted above—namely, feeling ‘uncomfortable in the
relevant distinctive way’—then our ordinary concept brooks no distinction
between the state of being in pain and the state of satisfying the criteria for being
in pain. So if, on the grounds reviewed, the descriptive phrase—‘the state of
satisfying the criteria for being in pain’—is rigid, then ‘pain’, too, must be rigid;
for only a rigid term can sustain a necessary identity with another rigid term.

So, hey presto! Kripke’s considerations, unless otherwise faulted, can indeed
be extended not merely to argue that pain is not a physical natural kind con-
cept but to distinguish the reference of ‘pain’ from any physical kind of state.

Notice that it has not been suggested that no type of physical state is
associated with pain, and thus to that extent characteristic of it. How could
philosophy establish that? The (none the less remarkable) conclusion is only
that no such type of state is pain—that our very concept of pain contains the
ingredients to prohibit any such identification.

1.8 The Boyd Objection

Let’s review the essential moves of the argument. Its essence is that whatever
physical identification of the state of pain is proposed, it will be prima-facie
conceivable that it might be empirically confounded. But the identification
must hold, if at all, then of necessity (if only for the rather complex reasons just
reviewed). Any claim of necessity may be defeated by conceiving—albeit
genuinely conceiving—of scenarios in which what it affirms to be necessary
does not obtain. So, absent some reason to think that the relevant scenarios
about pain somehow fail to portray what they purport to portray, we should
(defeasibly) conclude that no identification of pain with a physical state is ne-
cessary, or therefore true. Such a reason is available in the case of the apparently
conceivable scenarios in which water turns out not to be H₂O, since it has to
be ensured that the stuff of the conceived scenario really is water, rather than
something that merely possesses water’s surface symptoms. But no such difficulty
afflicts the case of pain and, say, C-fibre stimulation, since any ‘surface-sympto-
matic’ counterpart of pain is pain.
The crux, then—the point that is supposed to make all the difference—is that to conceive of a symptomatic counterpart of pain is to conceive a pain, whereas to conceive of a symptomatic counterpart of water is not, per se, to conceive of water. But on reflection, how can this be enough to make the difference? In order to conceive of an identity statement’s failing to hold, it suffices to conceive of one term in being while the other is not. Kripke thinks that we can conceive of a pain’s occurring without any C-fibre stimulation taking place, and vice versa. And sure, it seems prima-facie completely straightforward. Just imagine a situation in which you suffer pain, yet even the most sophisticated apparatus detects no C-fibre stimulation within you; or a situation in which, conversely, the C-fibre activity detectors go off the dial, while you lie relaxed in utter comfort. But the fact is that, before we can be entitled to take any prima-facie conception of the falsity of an identity statement at face value, it has to be that both its terms are resistant to the difficulty that Kripke makes for the attempt to conceive of water’s turning out not to be H₂O. Both the identified items must be such that we are entitled to regard the thought experiment as really engaging them, rather than mere symptomatic counterparts. Yet Kripke only considers pain. The other half of the imagined scenario—that there is, or is not, C-fibre stimulation involved—is not considered at all.

Now, situations in which C-fibres are stimulated would presumably form a natural kind. So, as in the case of water, the objection continues, there ought to be a distinction between instantiation of the kind and mere symptomatic imitation. That would be enough to create a problem for the suggestion that the necessity of ‘Pain is C-fibre stimulation’ might be defeated by conceiving of a situation in which C-fibres were stimulated yet no pain was felt. For while, presumably, the absence of pain, too, is a state that has only to seem to be in order to be, there ought, if there is a gap with water, to be a corresponding gap between conceiving of a genuine case of C-fibre stimulation and conceiving of a symptomatic counterpart. So that form of attempt to rebut the necessity of ‘Pain is C-fibre stimulation’ would be blocked by Kripke’s own move. Conversely, while the concept, situation in which C-fibres are not stimulated, is not a natural kind concept at all—since not a concept of a type of state of affairs with a unified underlying nature—essentially the same point may still seem good against the other relevant kind of counter-scenario: pain without C-fibre stimulation. Maybe I cannot fail to conceive of a pain when there is no C-fibre stimulation by dint of failing to conceive a genuine pain. But surely I may so fail by dint of failing to conceive of a genuine lack of C-fibre stimulation. For just as one does not conceive of a lack of water by conceiving of an absence of satisfaction of the indicator properties of water—the concept leaves provision for non-standard instances—so, it may be suggested, one will
not have conceived of an absence of C-fibre stimulation merely by conceiving of a lack of satisfaction of the indicators of C-fibre stimulation, whatever they are. It will not be enough merely to conceive of an absence, as assessed by whatever operational tests are appropriate, of the appearances of C-fibre stimulation.

The analogy is therefore apparently restored, and now a dilemma arises. If Kripke was successful in defeating the conceivability objection to the necessity of ‘Water is H₂O’, then his argument against physicalism collapses for want of a relevant disanalogy between ‘Pain is C-fibre stimulation’ and ‘Water is H₂O’. But if he wasn’t successful, then the whole apparatus of a posteriori necessities on which the argument against physicalism depends is put in jeopardy in any case.

Call this the Boyd objection. It is apt to seem a very damaging objection. But I do not think that it is persuasive at all. Let me try to explain why.

The failure of the attempted thought experiment in which water supposedly turns out not to be H₂O hinges on the claim that—when it seems that conception is possible—what turns out not to be H₂O is not distinguished from a mere water imitator, as it were. It is crucial to understand the source of this claim. In particular, it is not an instance of a general thesis about the limits of conceivability—there is no suggestion, for instance, that the conceiving faculty cannot encompass water per se at all, but only the appearances of it. The challenge to the author of the thought experiment is to specify how its subject is identified. He is not allowed just to reply, ‘As water’. Matters cannot be left there, or claims of counter-conceivability will become unnegotiable, since a proponent will have no explanatory obligations. An account is owing, accordingly, of what makes it water that is the subject of the imagined scenario—of why what is imagined should be regarded as water—and the (plausible) suggestion is that the thinker will prove to be relying on imagined satisfaction of the surface indicators. So it is perfectly fair to reply that, according to the view that is being opposed, that is insufficient to ensure the relevance of the thought experiment. But notice, to stress, that there is no claim here that mere conceiving can make nothing of the difference between genuine water and a symptomatic imitation—that the differences between them are, so to say, opaque to the conceiving faculty—so that any attempt to conceive of water would be bound to be no more than a conception of a display of indicators; on the contrary, to conceive of water, as opposed to conceiving of a symptomatic imitation, is to conceive of a purported natural kind—a substance with a certain underlying essence from which those symptoms characteristically flow. And still less is there a general claim to the effect that conceiving can make nothing of the difference between an appearance and reality—that its movement

12 See Boyd (1992).
is confined among imagined appearances. That would be a hopeless claim, for a reason to be noted in a minute.

Nevertheless, the Boyd objection would seem to be feeding on some such idea. To see this, reflect that in order for it to work, there has to be as good a reason to suppose that any apparently lucid conception of a situation in which there is, for example, pain but no C-fibre stimulation fails to represent a genuine lack of C-fibre stimulation as there is reason to think that the water/H₂O thought experiment succeeds in engaging no more than a symptomatic counterpart of water. Consider, then, what sort of thing might happen when someone—let her be an expert physiologist—who knows about C-fibres and their characteristic forms of activity tries to conceive a scenario in which they are inactive. Well, she’ll no doubt imagine certain tests and micro-physiological investigations turning out in a certain kind of way. And one conceptual gap that there will be is that between the appearance—the seeming that the tests and investigations turn out that way—and the reality—their actually doing so. So it would certainly serve the purpose of the Boyd objection if it were right that the conceiving faculty cannot cross this gap: that any thought experiment can engage no more than appearances. However, that’s a radical error (and no part of Kripke’s original point). The price of that contention would be that any apparent contingency whatever could be claimed with impunity to be a posteriori necessary—since any apparently perfectly lucidly conceived scenario in which it failed would be properly describable merely as one in which it appeared to fail. The link between conceivability and possibility may be subtle and qualified, but it is genuine—and to confine conceivability to appearances would be to sever it altogether. Thus, failing some independent reason for thinking her conception comes short, it should be granted that the physiologist really can conceive not just of an appearance but of the reality of the relevant tests’ militating against the hypothesis of C-fibre stimulation.

A proponent of the Boyd objection may try to regroup. Probably there will still be a gap—the tests may not be conclusive, or they may be liable to operational error. A situation that, by the most refined and painstaking tests that we have, must be classified as one in which no C-fibres are stimulated may still be misclassified as such. Agreed; but this gap will not suffice to drive the objection. Someone challenged to explain what makes it water that turns out not to be H₂O in his thought experiment doesn’t—from the point of view of one who accepts that water is a natural kind—have a good answer; for the answer involves mere satisfaction of the indicators. But if our physiologist is challenged to explain what makes it a lack of C-fibre stimulation that figures in her thought experiment, she has available the best possible answer: namely, its seeming and continuing to seem by the most refined tests that no C-fibre
stimulation is taking place. That is to be compared not to mere satisfaction of
the indicators of *water*, but to the most refined evidence of the presence of H₂O.
If someone thinks it is not a good enough answer, then forget about pain for a
minute and ask: what could conceiving of a situation in which it turned out
merely that there was no C-fibre stimulation consist in? But the conceivable of
*that*—on its own—wasn’t supposed to be in doubt.

To confirm that one who presses this objection is tacitly shifting the
goal-posts—is making a move quite different from Kripke’s—it ought to
suffice to reflect that it would presumably be conceded on all sides that, were
*water* to be supplanted by *schwater*, there should be no difficulty concerning
the conceivable of a variety of substances turning out to be schwater. It is
granted on all sides, in particular, that we can conceive of a symptomatic coun-
terpart of water turning out not to be H₂O. But the attempt to disqualify the
physiologist’s thought experiment along the lines just reviewed demands rules
of conceivable that would also disqualify that seemingly straightforward con-
ception. For how is it given that the conceived substance is not H₂O? Wouldn’t
it be merely that it turned out not to be such by the most refined extant chem-
ical tests, and wouldn’t they bear a merely defeasible relation to the fact?

In summary, what I have been saying boils down to the point that the Boyd
objection misses the distinction between primary and derivative natural kind
concepts, and the attendant different constraints involved in conceiving of
instances of them, H₂O and *C-fibre stimulation*, we may suppose, are primary
natural kind concepts. Accordingly, their reference is fixed not by advert-
ing to indicator properties, but directly, in the light of the explicit content of the
concepts themselves. These concepts are thus associated with no analogues of
the distinction between water and symptomatic counterparts of water, and
thought experiments in which they feature cannot be faulted for insensitivity to
such distinctions.¹³ A primary natural kind concept does provide, to be sure,
for two other types of distinction: between something’s appearing to be P and
its genuinely being so, and between something’s passing the most refined extant
tests for being P and its genuinely being so. But to insist that the Kripkean
thought experiments about pain and C-fibre stimulation are flawed by insen-
sitivity to the second of these distinctions is implicitly to disable a whole range
of perfectly valid conceivings, while to fault them for insensitivity to the first is
implicitly to confine conceiving to the realm of appearances, and thereby to
forfeit the connection between conceivable and modality altogether.

¹³ At least in so far as they relate to those two concepts, though there may of course be such
faults in connection with the involvement of other, derivative, natural kind concepts.
1.9 The McGinn Objection

A second very widely received objection to Kripke’s argument was first advanced by Colin McGinn (1977). Actually, McGinn’s objection presents itself as more of an accommodation; it does not abrade at all against the argument as I have so far presented it. Distinguish three claims: that pain is the concept of a physical natural kind, that pains form a physical natural kind, and that pains are physical. It is the third that is the essential thesis of physicalism. But it is at most a rejection of the first two that is supported by the considerations that we have reviewed. Even if pains are not a physical kind, each and every individual pain may nevertheless be a physical state. Pains may be token-identical with physical states, even if not type-identical.

That was McGinn’s point. Grant that any particular (neuro)physical type of event may coherently be conceived as dissociated from the occurrence of pains, and that this consideration can indeed be worked into a demonstration that pains are not a physical type. Still, the consideration seems powerless to engage the thought that the particular pain I am feeling now is token-identical with some aspect of the particular physical state I am in. Suppose the aspect in question is actually one of physical type F. So the supposition is that the pain I am now feeling is my being in the particular F-state that I am presently in. Plausibly, ‘The pain I am now feeling’ and ‘The F-state that I am currently in’ are both rigid designators. So the identity statement linking them, if true, holds of necessity. In the presence of the Counter-Conceivability Principle, it would therefore make trouble for the purported identity if I could coherently conceive of that statement’s failing to hold. But while I am granted a conception of how it could be that I was in pain without being in an F-state at all, that is not to conceive of my having this pain without being in an F-state, a fortiori without being in this F-state. For, it is very plausible, nothing in such a thought experiment engages the numerical identity of my present pain. No doubt I could conceive of having a qualitatively indistinguishable pain, under the very same circumstances, on the very same occasion. But if token–token physicalism is true, such conceiving may be regarded as portraying the pain I actually feel only if it involves nothing inconsistent with the actual physical identity of that pain. This is not to say, note, that conceiving is necessarily insensitive to the distinction between numerical identity and quantitative identity among token pains. It is to say, rather, that the content which may permissibly be assigned to a conceived scenario has to be sensitive to the essential characteristics of the items it involves.

If this is right, then the Kripkean argument may after all be consistent with physicalism. The strongest conclusion that it will be permissible to draw will
be, not that mental states—or at least, all mental states which are akin to pain in having a purely phenomenal essence—are not physical in nature, but that their physical identity is not that of a physical type.

1.10 The Explanatory Potential of Token-by-Token Physicalism

McGinn’s point seemingly provides a way of reconciling Kripke’s argument with the ontology of physicalism. But is the resulting form of physicalism worth having? In particular, to what extent can it preserve physicalism’s traditional advantages? Paramount among those was the prospect of the complete intelligibility of the world via the categories of physical science. That may now seem to be in jeopardy if the identity of, say, sensations with physical states goes merely token by token.

The line of concern, more fully, is as follows. To make scientific sense of the role of any state, property, or event in a world conceived as purely physical, it will be necessary to bring the item under concepts that render physical scientific laws applicable to it. The very generality that is of the essence of physical law, however, ensures that such laws will concern physical types. So if—while sensations, say, may all be physical, token by token—there are no physical properties with whose possession being in particular sensational states may generally—type by type—be identified, then in order to make sense of the role of such states in the physical world by the application to them of physical laws, it will be necessary to bring them under concepts that are at best, as it were, fortuitously coextensive with the concepts they fall under qua sensations. Suppose, for instance, that—as it happens—everyone in pain is in a state of C-fibre stimulation, and vice versa. Still, by the Kripkean argument, the state of being in pain is not the state of C-fibre stimulation. The concepts, pain and C-fibre stimulation, must still be reckoned to present different properties. Since the former property is not to be identified with any physical property, it follows that physical science can make no sense of—can give no scientific insight into—the coextensiveness that happens to obtain. And if it cannot do that, then it can never explain, for example, why aspirin eases a headache, no matter how convincing an account it has to offer of the effects of aspirin on C-fibre stimulation. By contrast, if the identity of pain and C-fibre stimulation goes type to type, then the scientific tractability of the latter just is that of the former. If headaches are a kind of C-fibre stimulation, then explaining the effects of aspirin on (that kind of) C-fibre stimulation is explaining its effects on headache.

In brief, if being in pain is the same thing as being in a state of C-fibre stimulation—if the identity goes type to type—then explaining aspirin’s effects
on C-fibre stimulation is explaining its effects on headache. Headaches become potentially scientifically tractable. But for token-by-token physicalism, it seems that the best that can be said of the types is that they are coextensive. Since this coextensiveness is then left explanatorily surd, we do not get explanations of the patterns of instantiation in one of the types merely by explaining the corresponding patterns in the instantiation of the other.

The concern is prima-facie compelling. But it makes an unsupported and crucial assumption. It assumes that if we may not identify pain with a physical type, then we are barred from identifying tokens of pain with tokens of a single physical type. This seems unwarranted. Consistently with rejecting the identity of pain, as a type of state, with the state of C-fibre stimulation, we might retain all explanatory advantages of that identification—if the empirical circumstances allow—by identifying each subject’s individual pain with a particular episode of C-fibre stimulation. Then, if best science were to find that, and to explain why, whenever a subject is in a state of C-fibre stimulation, its mitigation is indeed a normal effect of taking aspirin, the fact that C-fibre stimulation could not legitimately be identified with pain per se would be no obstacle to adapting this finding to the explanation of aspirin’s effect upon headaches. True, there would now be no way of converting the explanation of why aspirin mitigates C-fibre stimulation into an explanation of why aspirin relieves headache—but if each, or enough, individual headaches are token-identical with individual episodes of C-fibre stimulation, we still get an explanation of why aspirin relieves them. And isn’t that good enough?

If this is right, it points up something important: namely, how merely token-by-token identifications with the physical can always tap into the explanatory advantages that would have been secured by corresponding type-type identifications. Thus token-by-token theories need not, per se, involve any consequences regarding anomalousness. Sure, physical laws are essentially general, so are naturally formulated in terms of types of property, event, and state. Thus, in order to harness such laws to the explanation of what we presume to be the causes and effects of sensations, it might seem that we must find types of physical state for the sensations to be. And then, if such identifications are proscribed, it may seem as though some form of anomalousness, or scientific opacity, of sensation must be the upshot. But not so fast: the simple countervailing thought is that if a law connects one type of state with another, it thereby connects their tokens. So to treat of sensation in a fully intelligible but physical-scientific way, we do not need type-type identifications: it is enough that token sensations be token physical states of (some manageable range of) types that are tractable at the level of physical law.
1.11 A Discomfort about Supervenience

Nevertheless, I do not think that token-by-token physicalism offers a satisfactory accommodation with the Kripkean argument. One general difficulty with the proposal concerns supervenience. Most of us believe in some form of supervenience of the psychological upon the physical. Of course, supervenience relations come in many varieties. But I am referring just to the general idea that psychological differences demand physical ones, that had the psychological history of the world been different in any respect at all, then its physical history would have had to have differed too. Is this a rational belief? It is hard to be certain what exactly is its provenance. It is not empirical. Experience might suggest that many psychological differences tend to go with physical ones. But it could not suggest that they must so do. And if we seemed to alight upon psychological changes that went unreflected in any physical differences, we would insist that there must be physical differences all the same, though perhaps of a kind—arcane variations in brain state, maybe—of which we have no present conception.

This supervenience is certainly a rejectable thesis. It would, for instance, be rejected by Cartesian dualism. The supervenience requires that, as a matter of necessity, any change in a subject’s psychological condition must be attended by a change in her physical condition. This would be utterly incomprehensible if dualism were true: how could change in one ontological realm necessitate change in another? Dualism has it that the psychological and the physical are distinct existences. There is therefore no room for them to be linked as a matter of necessity. To that extent, the entrenchment of psychological-on-physical supervenience in our ordinary beliefs is indicative that our fundamental conception of the psychological is not Cartesian. But Wittgenstein—no Cartesian—also effectively rejected the supervenience (not there so termed, of course) in Zettel 608–10. And it is not clear that his stance is anything which we can readily confound by direct argument, either empirical or a priori.

If we have any justification for believing in psychological-on-physical supervenience, it would seem it must have less to do with what we can support by direct argument than with a sense of commitment to it flowing from our basic metaphysics of the psychological. But what metaphysics of the psychological would account for it? How could it be necessary that change in one range of states of affairs might necessitate change in another? The only possible answer,

---

14 This very general form of supervenience need not involve, of course, that the psychological supervenes upon the *internal* physiological states of the bearers of psychological properties. It is uncompromised by views that see certain psychological and semantic characteristics as broad.
it seems, is if they are not fundamentally of different categories but are, at bottom, states of one and the same sort, though presented in very different conceptual vocabularies. In short, if we do not believe that psychological states and processes fundamentally are physical ones, the claimed supervenience, far from being basic to our conception of the psychological, would be unintelligible, and the belief in it totally unmotivated.

Only some form of physicalism, it seems, can make sense of the supervenience. Thus it may appear that not only is a token–by–token physicalism consistent with the Kripkean argument, but that—if type–to–type physicalism is now ruled out—the supervenience of the psychological upon the physical actually demands it. However—and this is the advertised discomfort—it is not clear on reflection how token–by–token physicalism can actually accommodate the supervenience any better than can dualism. To be sure, if my present headache actually is some token physical condition of my central nervous system, then, had I not had the headache, I would have been in a different physical condition. But psychological–on–physical supervenience, in the form that is usually accepted, requires that a change in which psychological predicates may be truly applied to me requires change in a complete description of my physical state, where the latter description is precisely conceived as a compendium of the types of physical state I am in. Token–by–token physicalism simply cannot explain the validity of that principle. It requires, to be sure, that if I had not had the headache, I would have been in a different physical state; but it is open as to whether that different state could not still have been of the same type as that I am actually in. So it allows that the physical state I was in then could have been in every way type–indistinguishable from my actual physical state. Psychological–on–physical supervenience, as ordinarily understood, proscribes exactly that.

Note that this is not an objection to token–by–token physicalism’s ability to accommodate the Kripkean argument. The point is, rather, that the accommodation is available only if we either jettison the belief in supervenience altogether or are prepared to retain it in a setting in which we have neither direct argument for it nor any broader metaphysical justification.

1.12 Token–by–Token Physicalism and Rigidity

The most fundamental weakness in the token–by–token response to the Kripkean argument, however, is that, far from providing an accommodation with the conclusion, it proves on reflection to be committed to arguing with the premises. To the extent, then, that someone finds those premises to be well–motivated, she is precluded from responding in the way McGinn proposed.
Recall the dialectical situation at the crucial point. We noted earlier that the Kripkean argument was only able to advance beyond the stage of a claim about \textbf{pain} (that it is not a physical natural kind concept) to a claim about \textit{pain} (that states of pain are not a physical natural kind) courtesy of the assumption that ‘pain’ is a \textit{rigid designator}—that ‘is in pain’ ascribes the same state on all occasions of competent use. We canvassed argument for the assumption, which is indeed independently plausible in any case.\textsuperscript{15} But how can it be consistent with mere token-by-token physicalism about pain? For the distinctive thesis of the latter is precisely that while all pains are physical states, no single kind of physical state need be shared by subjects in pain.

More fully, the Kripkean argument appealed to the conceivability of a situation in which no specific type of (presumably) neural state is correlated with the occurrence of, say, migraine. Suppose that situation actually obtains. Still, for token-by-token physicalism, each and every migraine headache is identical to some specific token neural condition. Consider two such token headaches, say of sufferers $S_1$ and $S_2$, and the two token neural states, as it happens of utterly different physical types, say $F_1$ and $F_2$, with which those headaches are respectively token-identical. What physical states are the respective referents of the two uses of ‘headache’ in the two true claims ‘$S_1$ has a headache’ and ‘$S_2$ has a headache’? Since, for the physicalist, $S_1$’s headache just consists in her being in state $F_1$, and $S_2$’s headache consists in her being in state $F_2$, there seems no option but to allow that the two uses of ‘headache’ effect reference to these two different physical states. But then, for the token-by-token physicalist, ‘headache’ must be an expression referring \textit{flexibly} among those type physical states whose instantiation, on particular occasions, may constitute a subject’s having a headache—it cannot be a rigid designator of any particular such state. So in a physical world—where the only real types of state are physical—it cannot be a rigid designator at all.

Now McGinn’s intervention was pointless unless the Kripkean argument succeeds against the identification of pain with a physical type. It so succeeds only if ‘pain’ is rigid—otherwise the argument can be seen off by Australian Rules physicalism. But if ‘pain’ is rigid, what—physical—state can the token-by-token physicalist regard it as ascribing? The whole essence and being of her view is that no single type of physical state need constitute being in pain, that pains may be constituted by quite different types of physical state. Rather than providing a solution, then, token-by-token physicalism would seem to have no way of construing the data of the problem that McGinn introduced it to address.

\textsuperscript{15} In fact, all the same intuitions kick in as those Kripke appeals to for ‘Aristotle’: Could pain not have been pain? Could a different state have been the state of being in pain?
The point seems decisive. But a skirmish is possible. The objection could be finessed if we could take it for granted that there are, for instance, disjunctive states, or existentially general states. Equipped with such a repertoire, the physicalist might straightforwardly allow that the reference of 'headache', on every occasion of use, is indeed rigid—it is always to the disjunctive state of having $F_1$ or $F_2$ or $F_3$ or ..., where every type of physical state that can constitute a headache features as one disjunct. Or better, let the reference of 'headache' be to the state of being in some physical state which ... —and here the theorist plugs in her preferred account of the functional role of a headache (or of whatever the unifier is conceived as being). In contrast to type-to-type physicalism, individual pains could then be states of various physical types, united by their featuring in the relevant disjunction, or discharging the relevant functional role. But 'pain' could still be a rigid designator, denoting the same logically compound—disjunctive, or existential—state in every use.

But can physicalism regard such 'logically compound states' as in good standing? To begin with, one might wonder whether there is any satisfactory, principled specification of what might serve to unify the disjuncts of the putative disjunctive state, or to complete the specification of the putative existential one. And there is also a serious question whether such construals of the reference of 'pain' can make anything of another datum of the problem: the transparency of pains to their subjects. But, at least in the context of a physicalist ontology, the whole idea seems off target in a more basic way—specifically, by its illicitly imposing distinctions on to the domain of reference that make genuine sense only at the level of modes of presentation, concepts. When we disjoin or existentially generalize on names, the results—for instance, 'Tom or John was to blame' is better not be conceived as forms of expression involving reference to disjunctive or existentially general objects. There are no such objects. Why should it be different with states? Why should 'occupancy of some state with functional role $R$' be regarded as denoting a state at all? I do not deny the deflated sense of 'state' in which the nominalization of any significant predicative expression denotes a potential state. But physicalism is a serious ontological thesis, and one of its consequences is that the only genuine states there are, are physical. Suppose that amongst these are some with the functional role $R$. Clearly, 'occupancy of some state with functional role $R$' denotes none in particular of these, since it suffices to instantiate it that an object be in any of them. Equally, though, since these are the only states that have the functional role $R$, 'occupancy of some state with functional role $R$' does not denote some state outside that group. It follows that it does not denote a physical state at all, rigidly or otherwise. Since 'is in pain' and 'is in some state with functional role $R$' would be conceptually equivalent on the
present physicalist proposal, it would seem that the latter must repudiate the 
thesis that 'pain' is a designator at all, let alone a rigid one. But that again is to 
deny a datum of the problem.

2 Generalizations

2.1 Colour and Euthyphronic Concepts

Now to the case of colour. I choose it because there is a plausible case, so it 
seems to me, that colour concepts exhibit a germane kind of epistemic trans-
pparency. But whatever the facts about colour, it will emerge, if I am right, that 
any concept that is transparent in this way will lend itself to an argument of the 
Kripkean kind.

To fix ideas, let me rapidly rehearse the idea of judgement-dependent concepts 
which I have discussed elsewhere.¹⁶ Let a provisional equation be an instance of 
the following schema:

\[ C(S, x) \rightarrow (F(x) \leftrightarrow \text{it seems to } S \text{ that } F(x)) \]

That is, if conditions C are met by a subject S and an item x, then x is F if 
and only if it seems to S that it is. A concept, F, is Euthyphronic if such a provi-
sional equation can be written for it meeting each of the following four condi-
tions:

(i) The provisional equation is true a priori, as a matter of conceptual 
necessity.
(ii) The conditions, C, are specified in specific, substantial terms.
(iii) The satisfaction of the C-conditions is a matter that is independent of the 
details of F's actual extension.
(iv) The provisional equation is primitively a priori—it admits of no proof from 
ulterior principles concerning F of such a kind as to vindicate the idea 
that the C-conditions merely enable a subject to keep infallible track of an 
independently determined extension.

These clauses were proposed as one way of explicating the intuitive idea of a 
concept whose extension is constitutively sensitive to those of our verdicts that are 
delivered under what we conceive as the very best possible circumstances, 
rather than merely reflected by such verdicts. Whether that idea can indeed be 
so captured, whether all four conditions are necessary to capture it, and how

¹⁶ See the appendix to ch. 3 of Wright (1992).
exactly they might need to be modified or elaborated in order to succeed, are matters beyond the scope of this chapter. However, notice that pain plausibly sustains a very simple provisional equation meeting at least three of the specified conditions,17 namely:

S understands what pain is and is cognitively lucid \( \rightarrow \) (S is in pain \( \leftrightarrow \) it seems to S that she is in pain).

Moreover—at the price, perhaps, of a surprising degree of complication—a case can be made for thinking that the same is true at least of central colour concepts, like red.

The case requires a much more detailed discussion than I can digress to offer here.18 However, to sample its flavour, consider whether you think you have any clear concept of how the redness of a red surface could escape your judgement, or how its seeming red could be deceptive, if (i) the surface is in full view, and (ii) in normal daylight, (iii) relatively stationary (i.e., stationary or slow-moving relative to you the observer), and (iv) quite close by; and (v) if you know which object is in question, (vi) observe it attentively, (vii) are possessed of actually typical visual equipment, and (viii) are free of spots before the eyes, after-images, and so on, and (ix) are otherwise cognitively lucid, and (x) are competent with the concept red. You can add, if you like, that (xi) the surface be presented against a matt black background,19 and (xii) that you—the judging subject—be free of doubt about the satisfaction of any of these conditions. Anyway, the thought of one who conjectures that red is Euthyphronic is that it is possible in this way to construct a list of substantial conditions whose satisfaction ensures a priori that a presented item is red only if it seems so to an observer, and that the result will be a primitive truth about red: that the extension of the concept is, in such a fashion, constrained of necessity to be sensitive to our judgements under the elaborated ideal conditions.

Someone might wonder how so relatively complex a claim, which in the nature of the case lies beyond any rigorous proof, can possibly rank as a priori. But that shouldn’t be a sticking point. As a rough parallel, consider Church’s thesis, that every effectively calculable function is generally recursive, and vice versa. Effective calculability is an intuitive notion; general recursiveness is a

17 The fourth, anti-tracking condition is interestingly controversial in this case. John McDowell’s (1994) view, that sensations are essentially conceptualized modes of experience, is in effect the view that it is satisfied—that there is no brute phenomenal happening, mere pain, whose occurrence is indifferent to the conceptual resources of a sufferer and of which possession of pain merely enables him to keep track. But the fourth condition is in any case strictly inessential to the considerations to follow.

18 For further discussion, see Wright (1989).

19 Or maybe viewed through an apparatus—a tube, say—that occludes any background.
mathematically precise one. The thesis is precisely an attempt to give a mathematically exact characterization of something pre-formal. In the nature of the case, it therefore admits of no conclusive formal proof. Yet, if it is true, it is true purely as a reflection of the character of the concepts involved, and to the extent that it can be supported by conceptual reflection—for instance, by the striking convergences of other attempted formal characterizations of effective calculability, and a failure to find counter-examples—to that extent, it is supported a priori. It is similar with the provisional equation for red. Our concept of the variety of ways in which the redness of an object might in principle be masked by how it seems, or in which how it seems might be deceptive, ought to allow of correct circumscription, just as the concept of effective calculability ought. If we alight upon such a circumscription, it will certainly be too complicated to enable its truth to be recognizable immediately, just by the light cast by the analytic understanding, as it were; and there is no basis on which its truth might be recognized inferentially. As with Church’s thesis, its a priori correctness, if it is correct, will ultimately be supportable only defeasibly, by the failure of hard reflection to find it wanting.

There is a lot more to say. But our concern here is not with the justification of claims of judgement-dependence, but with the implications if certain concepts are judgement-dependent in accordance with the template described. Now the key point in the Kripkean argument was the claimed counter-conceivability of any physicalistic identification of pain. And the key assumption for that claim was that there is no coherent distinction to be drawn between a state’s seeming to its subject, S, to be one of pain and its actually being so. But that equation is not of course unconditional: it is, again, the impressions of a cognitively lucid subject, who fully understands pain, that there is no distinguishing from the fact of pain. And it is the standing prima-facie conceivability of a separation between such impressions and any specific type of physical (neural) state that constitutes the crucial point in the argument: what is claimed to be counter-conceivable, in other words, is precisely what results from the provisional equation above when the left-hand side of its biconditional consequent is replaced by a clause ascribing to S any particular physical state taken to be identical with pain:

\[ S \text{ understands what pain is and is cognitively lucid} \Rightarrow (S \text{ is in physical state } O \leftrightarrow \text{it seems to } S \text{ that she is in pain}) . \]

The crux in the Kripkean argument is thus, in effect, the contention that any provisional equation instantiating that schema on ‘O’ is counter-conceivable; but that one such would have to hold of necessity if pain were indeed a (type of) physical state.
The possibility of generalization to \textbf{red} is accordingly evident. Assume that we can indeed formulate a (conceptually) necessarily true provisional equation for \textbf{red} meeting the outlined conditions:

(1) \(C(S,x) \rightarrow (x \text{ is red} \iff \text{it seems to S that } x \text{ is red})\).

Assume also, for \textit{reductio}, that \textbf{red} is a derivative physical natural kind concept. Since there actually are red things, there will therefore be some presumably microphysical property, \(O\), such that

(2) Redness is \(O\)

is likewise necessary. We may take it\(^{20}\) that the necessity issuing a priori from concepts and that issuing a posteriori from essences are the very same, absolute metaphysical necessity, differing in the grounds for ascribing them to particular claims, but not in what is ascribed. That being so, since (2) entails that everything red is \(O\), and vice versa, we may infer that

(3) \(C(S,x) \rightarrow (x \text{ is } O \iff \text{it seems to S that } x \text{ is red})\)

is likewise necessary. But, for reasons exactly analogous to those which applied to \textbf{pain}, (3) ought not, it appears, to be regarded as necessary. For no matter what the detail of the physical kind, \(O\), it seems that it will be readily conceivable that the \(C\)-conditions are met—we are standing out of doors and out of shadow at noon on a typical cloudy summer’s day, staring at a stationary object quite close by; we are blessed with statistically normal visual equipment, we are attentive, and so on—and that we judge that the relevant object, which looks manifestly tomato-red, is red, and yet microphysical investigation discloses that it is \textit{not} \(O\). (Of course, there is at this point a debate to be had about the status of our apparently conceiving that conditions are indeed \(C\) and that \(x\) is indeed not \(O\); but this will merely recapitulate considerations which we went through in connection with the Boyd objection.) In general, the best case for supposing that, no matter what \(O\) might be, (3) is counter-conceivable promises to be exactly parallel to that for the counter-conceivability of ‘Pain is \(C\)-fibre stimulation’. The only difference is the relative complexity of the \(C\)-conditions.

In the first instance, this will be a conclusion not about redness but about the concept, \textbf{red}, and, as previously, a prima-facie accommodation with the argument will be available if the physicalist about colour is willing to maintain that (2), although true, is indeed not necessary, because ‘redness’ is a flexible

\(^{20}\) As of course, implicitly, did Kripke. For it is presumably a \textit{conceptual} necessity that there is no distinction between being in pain and being in a state epistemically indistinguishable from pain.
designator among physical natural kinds, equivalent in sense, perhaps, to the (unrigidified) description, ‘the physical state indicated by the property of looking \textit{red} under C-conditions’. Now recall that when we considered the corresponding response in the dialectic concerning pain and ‘pain’, we confronted the physicalist with the designator ‘the state of satisfying the criteria for being in pain’, which is plausibly both rigid and necessarily co-referential with ‘pain’. Since only a rigid designator can be necessarily co-referential with a rigid designator, it followed that ‘pain’ is rigid too, and that the physicalist’s flexible ersatz is false to the ordinary understanding of the term. Matters can now proceed in an analogous fashion with ‘redness’. We form, for instance, the description, ‘that state that, necessarily, anything is in that looks red under C-conditions’ (or—for an arbitrary judgement-dependent concept, F—‘that state that, necessarily, anything is in that seems F under C-conditions’). Then, provided the provisional equation is necessary, the intuition is strong that ‘redness’ and ‘that state which, necessarily, anything is in that looks red under C-conditions’ are necessarily coextensive; and the latter must, of course, be rigid if it refers at all, since otherwise there is no state such that necessarily anything is in it that looks red under C-conditions. It follows that ‘redness’ is rigid, that the property it denotes is consequently identical with no physical type—and that all this is called for by our concept \textit{red}: our concept of the kind of property redness is, if \textit{red} is Euthyphronic. As before, the question may be raised as to whether some form of token-by-token physicalism can save the day—and specifically whether it can accommodate the rigidity of the terms in question (for if that rigidity is denied, then the Australian Rules version already has all the resources the physicalist needs to see the argument off). And, for reasons analogous to those recently canvassed, we may well feel that it cannot.

2.2 \textit{Summary of the Recipe}

Let us take stock. The suggestion is that Kripke’s argument is apt to generalize to any concept, \( F \), of which it holds a priori, as a matter of conceptual necessity, that under certain substantially specifiable, conceivable conditions, it will seem to a thinker that the concept is instantiated just in case it is. For \textit{pain}—as, presumably, for sensation concepts generally—the conditions in question are very simple: they are merely that the thinker grasps the concept in question and is appropriately cognitively lucid. Other cases will not be so simple: if \textit{red} indeed comes in this category, its C-conditions will be complex. Still, the thesis is a going concern that our concepts of colour, and of Locke’s secondary qualities generally, do have the requisite kind of transparency: that even if it is harder than
philosophers once supposed to say under what circumstances it is equivalent to an object's being red that it look red, it remains that, with care, one can produce a list of conditions such that nothing counts as an explanation of how, under those conditions, an illusion of colour could occur; and that the same holds for sounds, tastes, and secondary qualities generally. To have constructed such a list for a given concept F—a list for which no one can produce any prima-facie acceptable account of how under the specified conditions it could seem to thinkers that F was instantiated although it was not—will be to have an a priori, albeit defeasible, case for regarding the relevant provisional equation as holding as a matter of conceptual necessity. That, as illustrated, will then suffice to set a form of the Kripkean argument in train.

It will do so by supplying each of two needed premisses. The first premiss, the necessitated provisional equation

Necessarily: for any S and x under C-conditions, x is F iff x seems F to S

is supplied directly. The second premiss, the necessity of any true identity of the form

F is K,

where 'K' rigidly denotes a physical kind, requires that 'F' too is a rigid designator. But, as we saw, the provisional equation's holding of necessity arguably ensures that as well. It does so provided we may abstract from it to infer that being F is being in that state P such that, necessarily, any object is in P that seems F to a thinker operating under C-conditions. For again, if 'F' were flexible, there would be no single state of which it would be necessary that any object—in whatever world—seeming F to a C-conditioned thinker would occupy it; rather, being F might be being in P₁ in w₁, and being in P₂ in w₂, and so on, and of no Pₖ would it be true that in all worlds and for any thinker operating under C-conditions, x's seeming F would be necessary and sufficient for its being in Pₖ.

The two premisses collectively entail that 'F' may be substituted by 'K' in the first to generate a further necessary truth: that for any S and x in the appropriate C-conditions, x is K iff x seems F to S. But the necessity of this consequence is likely to be in difficulty with the Counter-Conceivability Principle, provided conditions C are ones that we can conceive to obtain—which, at least if the kind of list sketched for red is any guide, there seems every reason to

---

33 Recall Wilfrid Sellars: 'But what, then, are we to make of the necessary truth—and it is, of course, a necessary truth—that x is red + x would look red to standard observers in standard conditions?'—from 'Empiricism and the Philosophy of Mind' reprinted in Sellars (1963: 142). Cf. Colin McGinn: 'It is a conceptual truth that red things typically look red' (1983: 11).
anticipate. Moreover, since the list is likely to proceed in quite general terms of ideallization—explaining wherein consist good observation conditions, competent observers, and so on—without any reference to how in particular things seem to a judging subject or how they actually physically are in other respects, there seems no reason why there should ever be any tension in adjoining to an imaginary scenario in which the C-conditions hold the further detail that the thinkers involved have the impression that x is F; or to that second scenario in turn the additional detail of x’s turning out under examination not to be K.

In general, the Kripkean argument will thus extend to all concepts for which an appropriate conceptually necessary provisional biconditional can be constructed, the obtaining of whose C-conditions is prima-facie conceivable, and where such a scenario may prima-facie coherently be augmented by its seeming to a thinker under the C-conditions in question that a given item instantiates (or fails to instantiate) the concept in question, while that item simultaneously lacks (or has) any particular candidate to be the physical property presented by the concept F. To be sure, in view of the prima-facie (defeasible) status of the relevant conceivings, the argument is best viewed as a challenge to physicalism, rather than a purported refutation. An undefeated impression that one can simultaneously conceive of a thinker’s meeting the C-conditions and, say, judging that x is F while x lacks candidate physical property, O, is, by the Counter-Conceivability Principle and assuming the rigidity of the relevant terms, an undefeated challenge to the physicalist identification. But views that confront undefeated challenges ought not, aeternis paribus, to be believed.

Kripke’s argument poses such a challenge on behalf of concepts of sensation. The challenge extends, I have suggested, to concepts of colour and, one would expect, to secondary-quality concepts generally. However, we should also note that, in moving directly to the case of secondary-quality concepts, we have passed over the possibly more straightforward, intermediate case of concepts of public subjective appearance—looks red, feels hot, seems peppermint—on which we tend to fall back in cases where there is a potential for a mistake or an illusion about a secondary quality for reasons other than the idiosyncrasies of individuals, and which one would expect to be associated with a much simpler set of C-conditions than those for secondary-quality concepts proper. The point is also salient that even if our actual secondary-quality concepts prove recalcitrant, on closer inspection, to the construction of the relevant kinds of provisional equations, that would seem to be the merest good luck as far as physicalism is concerned. For if the primary function of secondary-quality concepts is to provide means whereby we can record aspects of the world of
common subjective appearance, it is not clear that that role would be compromised in any essential way had they been so fashioned as to sustain the requirements of the argument after all. Even if red is not transparent in the necessary way, there could be a concept, shred, of which it was an a priori essential characteristic that the C-conditions I listed earlier were exactly right. How would the intelligibility and utility of ordinary discourse about colours be damaged if shred supplanted red, and our other colour concepts differed in parallel?

2.3 Other Cases

The large metaphysical dilemma with which we started mentioned semantics, psychology, and value. One might now investigate the prospects for detailed provisional equations for those discourses’ characteristic concepts which satisfy the first three conditions of Euthyphronism and thereby provide the basics for the Kripkean argument. But a less demanding form of epistemic transparency may suffice. It is arguably part of our concept of ordinary intentional psychological states that they are standardly manifest to their subjects, who are in turn standardly authoritative about them. It is no mere empirical truth, that is to say, that normally a subject will know what she believes, hopes, wants, intends, and so on, and that her opinions on such matters will be right. At least in cases where one would expect self-knowledge to be non-inferential (involving no conscious self-interpretation), it belongs, plausibly, to the very concept of a subject’s being an intentional subject at all that her impressions of her intentional states are generally reliable, and that those states do not generally escape her. We can, to be sure, break the conditional in either direction: sometimes it’s fruitful to think of subjects as self-deceived or as self-unaware. But these don’t count as good options unless an interpreter is able to back them up with independently attested details about the subject’s frame of mind and an explanation of why the circumstances might have been conducive to the mismatch in question. Absent such an account, the rule is that intentional facts should be reckoned to march in step with the subject’s impression of them.

This point, if granted, is enough for a variant form of the Kripkean challenge, concentrating on what we might term scenarios of multiple exception. Provided it is accepted that expressions denoting types of intentional state do so rigidly, there will be a standing obstacle to any proposed identification of such states with physical types posed by the apparently lucid conceivable-ability of a world in which it is the normal case that subjects who have the impression that they are in particular such states fail to instantiate the appropriate physical types, while no considerations are to hand—or indeed,
emerge—that would allow an interpreter to regard them as self-deceived or self-unaware.

With semantics, the trick is to revert more closely to the Euthyphronic template, but to multiply the judges. It may be arguable whether communally pervasive ignorance of meanings is possible: whether our concept of the meaning of an expression allows for the possibility that sentences which are, for instance, too complex or convoluted to be reliably parsed by ordinary speakers might nevertheless have determinate meanings. But it seems that we can a priori exclude any possibility of communally pervasive error: any possibility that all or a large majority of, by normal criteria, competent speakers should mistake the meaning of an expression—that they should take it, and recognize each other as taking it, as meaning one thing when in fact it means another. Rather, if normal speakers are mutually recognized as attaching a particular meaning to an expression, then that is what it means, punct. So again, assuming that expressions of the form ‘means that’ P are rigid—that they ascribe the same property in speaking of expressions in any world—there will be a standing challenge to the necessity, and hence the truth, of any proposed identification of such semantic properties with (presumably relational) physical types: the challenge will be posed by the apparently lucid conceivability of a scenario in which the members of a seemingly smooth-running speech community exceptionlessly understand themselves and each other as meaning that P by an expression that lacks whatever natural properties may have been proposed as being those in which meaning that P consists.

The case of moral concepts, finally, is more qualified. There are many, usually theistic, conceptions of morality that view human moral thinking as essentially imperfect; and there are others—for instance, some forms of utilitarianism—that view it as almost always limited by ignorance of relevant, non-moral facts. But one would expect a form of the generalized Kripkean challenge to apply on many humanistic or ‘ideal-observer’ accounts of the moral, according to which there is no gap between moral quality and the assessment of it offered by the most rational, fortunately situated judge. For again, it looks as though there should be no barrier to conceiving of a situation in which, in the context of any particular proposed naturalistic account of some moral quality, M, a thinker rates a situation as, for example, possessing M, where the relevant naturalistic features are missing, and where there is nothing to impugn the moral credentials of the thinker in question. Assuming the rigidity of moral predicates—that expressions for moral qualities always ascribe the same qualities when used in speaking of hypothetical scenarios—there will then be the same standing difficulty, at least on such transparent conceptions of morality, to the location of moral qualities in the world of modern naturalism.
3 Denouement

3.1 The Counter-Conceivability Principle Again

We have now reviewed a range of variations upon the theme of epistemic transparency, seen why epistemically transparent concepts as a species may offer a Kripkean challenge to naturalist reductions, and noted that the distinctive concepts of many of the subject-matters with which modern naturalism abrades seem to be in the frame. Of course, the challenge is potentially reversible. Let it have been shown that naturalism is indeed in difficulties when it tries to accommodate the concepts in question by tending to construe them, after the fashion of water outlined at the start, as concepts of underlying physical kinds. That may be taken as bad news for naturalism. But it could be construed instead as bad news for the belief that the concepts in question are of things to be met with in the real—natural—world. Of course, that belief will prove pretty resilient when the concepts in question are concepts of one’s own sensations! But the dialectical point remains, that there is the option of error theory: it is naturalism as a descriptive, rather than as a revisionary, metaphysics which is put under pressure by the Kripkean argument.

But is it really put under pressure at all? An appreciation of the potential generality of the Kripkean argument, far from deepening one’s sense of a crisis for naturalism, is more likely to reinforce the impression that the conclusion is much too easily reached. My final point will be to corroborate that impression.

The Counter-Conceivability Principle says that a posteriori necessities, no less than a priori ones, are defeasible by lucid counter-conception: that if one can construct what appears to be a genuinely coherent scenario in which a putative necessity fails to hold, then, unless or until some shortcoming is disclosed in that conception, the claim of necessity should be regarded as defeated. Now, Kripke recognized straightaway that with a posteriori necessities in general, a prima-facie counter-conception will always be available—since one has only, apparently lucidly, to conceive of the relevant kind of a posteriori inquiry as turning out differently. Thus one may apparently lucidly conceive of an investigation into the chemistry of water that finds that it is XYZ; or apparently lucidly conceive of an astronomical investigation that discloses that the evening star is actually Jupiter; or apparently lucidly conceive of an investigation into the composition of the Macintosh computer before me whereby it turns out to be made entirely of materials of vegetable origin. It cannot plausibly be denied that some sort of coherent scenario is involved in such cases. But, assuming that the examples do involve genuine necessities of constitution and identity, the scenarios in question are not, by one who—like Kripke—accepts
the Counter-Conceivability Principle, allowed to count as genuine counter-conceptions. Rather, their claim to concern the items they are supposed to concern is found wanting. It is not water but something that presents itself as water—a mere symptomatic counterpart—that one can lucidly conceive turning out not to be H₂O; it is not the evening star, but a different, though similar-looking, body, imagined as occupying the same place in the evening sky, that one can conceive turning out to be Jupiter; it is not this computer, but one just like it, that one can lucidly conceive as turning out to be of vegetable materials.

What stopped this way of protecting the Counter-Conceivability Principle being totally devoid of interest—what stopped it being simple ‘monster-barring’—was that, precisely because the move is unavailable in the case of epistemically transparent concepts, the principle is allowed to retain some teeth. We do not have carte blanche. The claim to have conceived a scenario running counter to a putative necessity involving an epistemically transparent concept cannot be dismissed on the ground that the scenario fails to reflect the difference between a genuine instance of that concept and a mere symptomatic counterpart—for there is no such difference to reflect.

This was the ur-thought in the Kripkean argument. But it is sufficient for the purpose only if, where a posteriori necessities are concerned, there is no other way in which a lucid putative counter-conception may fall short than by failing to engage the distinction between an item and surface counterparts of it. For suppose there was another way. Then the apparently lucid conception of a scenario in which pain occurred without C-fibre stimulation, while it could not be dismissed as involving nothing distinguished from a mere surface counterpart of pain, might yet come up short in this other way. However, it is clear on reflection that we must make room for other possible kinds of shortcoming in purported counter-conceptions, however vivid and lucid-seeming they may be.

Here is an example. Suppose—Kripke would agree—that I am essentially a human being, and that it is an essential characteristic of human beings to have their actual biological origins. So it is an essential characteristic of mine to be the child of my actual parents. Still, I can, it seems, lucidly conceive of my not having had those parents but others, or even—like Superman—of my having originated in a different world, of a different race, and having been visited on earth from afar and brought up as their own by the people whom I take to be my biological father and mother. I can, it seems, lucidly imagine my finding all this out tomorrow. And it is, prima facie, every bit as coherent a scenario as those involved in the water, evening star, and computer cases. But it cannot be dismissed, in the way that they were, on the ground that it fails to be sensitive to
the distinction between myself and a mere epistemic counterpart, a mere ‘fool’s self’, as it were, sharing the surface features by which I identify myself but differing in essence. It cannot be so dismissed because I don’t, in the relevant fashion, identify myself by features, surface or otherwise, at all. The point is of a piece with Hume’s observation that, in awareness of a psychological state as one’s own, one is not presented as an object to oneself. When I conceive some simple counterfactual contingency—say, my being right at this moment in the Grand Canyon—I do not imagine someone’s being there who presents themself, on the surface, as being me. Rather, I simply imagine my having relevant kinds of experience—imagine, that is to say, the relevant kinds of experience from my first-personal point of view. No mode of presentation of the self need feature in the exercise before it can count as presenting a scenario in which I am in the Grand Canyon; a fortiori, no superficial mode of presentation, open to instantiation by someone other than myself.

The apparently lucid conceivability of the Superman scenario turning out to be the truth is a like case. Since it need involve no play with a mode of presentation of the self, it need not be open to any charge of insensitivity to the distinction between the self and a surface counterpart (whatever that would be). And of course, it would not help to try applying this complaint to other items in the conceived scenario—for instance, to the people, looking and behaving very much like my actual parents, who are now conceived as having fostered me instead. For even if the scenario is taken to be insensitive to the distinction between those who actually reared me and mere surface counterparts of them, it remains part of it that I was not born to the human race at all. Nevertheless, if I was indeed born to those I take to be my actual parents and the Counter-Conceivability Principle is correct, I cannot genuinely conceive of the Superman scenario. So how am I to describe the content of the scenario which I do seem to be able to entertain, in as great a degree of fanciful detail, moreover, as may be wished?

Next consider a different kind of example altogether. We can rest assured, I suppose, that Andrew Wiles really has proved Fermat’s Last Theorem, which therefore holds good as a matter of conceptual necessity. But we can imagine a sceptic about the result who flatters himself that he can still conceive of finding counter-examples to the theorem, and of finding mistakes in Wiles’s proof. Of course, there will be limits on the detail of these ‘conceivings’, or the sceptic would be thought—experimentally finding real counter-examples and mistakes. Still, we should not deny that he could be conceiving something, and doing so moreover—subject only to the preceding point—in as vivid and detailed a way as could be wished. The last diagnosis we should propose, however, is that his conceivings are insensitive to the distinction between finding counter-examples
to Fermat's theorem and finding counter-examples to an *epistemic counterpart* of it—or to the distinction between finding a mistake in Wiles's reasoning and finding a mistake in an *epistemic counterpart* of that. What could that mean? Of course, we can imagine someone with only a hazy idea of the theorem or the proof. But our sceptic may be perfectly clear about both—an able (but curmudgeonly) mathematician.

If we are to retain the Counter-Conceivability Principle, then we must provide house-room for these cases. There has to be a category of conceivings that fall short of being genuine counter-conceptions to a given proposition, not because their detail fails to be sensitive to the distinction between items that the proposition is about and ‘fool’s’ equivalents of them, but because it is insensitive to another distinction: that between genuinely conceiving of a scenario in which P fails to obtain and conceiving, rather, of what it would be like if, *per impossibile*, P were (found to be) false. The latter is what the curmudgeonly mathematician does. It is what I do when I conceive, in as much detail as you like, of my originating of a different race, elsewhere in the galaxy (or perhaps, following Descartes, of surviving my bodily death.) If time travel is metaphysically impossible, it’s what anyone does who imagines himself as a Time Lord, wandering in the fourth dimension. For a large class of impossibilities, there are still determinate ways things would seem if they obtained.³³ If, *per impossibile*, Wiles’s proof is flawed, and there are counter-examples to Fermat’s theorem, we know how things would seem if those circumstances came to light.

### 3.2 The Failure of the Kripkean Argument

To admit such a category of conceivings is not, let me stress, to make a concession inconsistent with the Counter-Conceivability Principle. It can remain that a prima-facie lucid conception of a scenario in which not-P holds is a standing, though defeasible, objection to any claim of the necessity of P. But if even a vivid and detailed scenario is to motivate such an objection, then it needs to be able to defend against not one, but two, charges of potential insufficiency. The first is the original Kripkean charge that nothing in its detail distinguishes its objects from surface counterparts—‘fool’s’ equivalents—of the items that it is supposed to concern; and the second is the charge that it allows of description as a scenario, merely, of how some things would or might be—for instance, what kinds of things would be experienced—if the proposition in question were false. If the scenario can be done full justice by a description of the latter

---

³³ This is one reason why semantical treatments of subjunctive conditionals that hold all with impossible antecedents to be true are unfortunate.
kind, then it will not be done less than justice if the description is modified by
the insertion of the words, ‘per impossibile’, and in that case it fails to constitute a
genuine counter-conception at all.

Both kinds of insufficiency afflict the scenario in which water purportedly
turns out to be something other than H₂O. But the price we pay for
retaining the Counter-Conceivability Principle is that the second kind of
insufficiency—per impossibile insufficiency, as it were—must be reckoned to
afflict any seemingly lucidly conceived scenario that appears to jar with a
necessity, whether a posteriori or not. So if P is a proposition that is known to
be necessary if true (and, correspondingly, impossible if false), then in order to
determine whether we have constructed a genuine counter-conception to P,
as opposed merely to a lucid scenario of how in certain respects things would
be if, per impossibile, P did not obtain, we need first to know whether P is true. If
P is true, then no matter how intricate and coherent, the scenario can embody
no more than a per impossibile counter-conception; if P is false, then there need
be no barrier to its description as a valid counter-conception. But either way,
the distinction turns, in the end, on matters beyond the phenomenology of
conceiving.

The effect is that, although we save the letter of the Counter-Conceivability
Principle, and although—for all I have said—it can continue as a defeasible
operational constraint on the ascription of necessity in cases where contingency
is an epistemic possibility, it provides no practical controls at all on the ascription
of necessity in cases where necessity would follow from truth—as is the situa-
tion of all potential necessities a posteriori. The consequence is that the appar-
ent counter-conceivability of physicalistic identifications of the instances of
epistemically transparent concepts is of no modal significance whatever.
Rather, the truth-values and hence—on the assumptions of the argument—
the modal status of the identities of the relevant kind must be settled independ-
ently, before we can know how properly to describe the scenarios in question.

So, after all, the metaphysical prospects for naturalism cannot be dashed
purely by creative exercises of the conceiving faculty, in the extraordinary
fashion that the Kripkean argument seemed to promise. But I don’t think that
we (most of us) ever really believed they could.

REFERENCES

Not Entail’, in Ned Block (ed.), Readings in the Philosophy of Psychology, i


