Does Conceivability Imply Possibility?:
Relieving Kripke's Worries about Mind-Body Identity

Abstract: Kripke argues that mental states are not identical to neurological states, for identities hold necessarily, and we can conceive of the alleged identity failing to hold. This argument assumes that conceivability is a reliable guide to possibility; but I offer a sketch of an account of possibility that suggests otherwise. Conceivability is often a reliable guide to possibility, but not always; I give a plausible criterion that (1) preserves the reliability of our intuitions about possibility over a broad range of cases and (2) explains why our ability to conceive of non-identity between mental and neurological states does not supply any evidence for the genuine possibility of such a non-identity. I thus claim to have removed this particular barrier to various physicalist theories of mind.

Does Conceivability Imply Possibility?
Relieving Kripke's Worries about Mind-Body Identity[1]

Saul Kripke has revived a Cartesian argument against the identity of mental states with physical states. His argument, which I will review in section I, hinges on a premise about the relation between conceivability and possibility, namely, that conceivability is a good guide to possibility. In section II, I will consider some broad intuitions about the nature of possibility and sketch how we might systematize these intuitions. With this sketch in hand, I will return to Kripke's argument in section III and attempt to show that the crucial premise about conceivability and possibility is false, though a slightly weaker claim is true. I suggest that conceivability is a good guide to possibility only in certain cases, and I explain how to recognize such cases. But Kripke's argument against the identity of mental states and physical states fails if only this weaker claim is true. Thus I hope to have removed this particular barrier to identity theorists and other physicalists.

I

Suppose we have some rigid designators *a* and *b*. (Recall that a rigid designator is a term that picks out the same object in every possible world.) Then, as Kripke's *Naming and Necessity* famously shows, the statement "*a* is identical to *b*" is necessarily true if true at all. For example, 'Hesperus' refers to a particular thing, and 'Phosphorous' to the very same thing, and naturally in any possible world this thing is identical to itself. So the statement 'Hesperus is Phosphorous' is necessarily true, as it holds in all possible worlds. If, on the other hand, we have some non-rigid designators *c* and *d* – where a non-rigid designator is a term that does not pick out the same object in every possible world – then the statement "*c* is identical to *d*" may not be a necessary truth. (But the *object(s)* referred to in this world by *c* and *d* will still be necessarily identical, if identical at all.)[i] Further, there is a simple test for rigidity. If the claim '*x* could have not been *x*' is true, then *x* is a non-rigid designator; otherwise it is a rigid designator. For instance: 'the inventor of the bifocals could have not been the inventor of the bifocals' is true, because surely Benjamin Franklin could have failed to invent bifocals, so 'the inventor of the bifocals' is a non-rigid designator. But 'Nixon could have not been Nixon' is false, so 'Nixon' designates rigidly.

Apply these observations to the claims of mind-body identity theorists like Armstrong. These theorists want to say something like this: 'Any mental state is identical to some neurological state.' Mental states such as pain are supposed to be identical to, say, C-fiber firings. This identity, Kripke notes, cannot be contingent. For when we apply our test for rigidity and ask, "Could pain have not been pain?" the answer is *no*. Likewise, Kripke supposes, for C-fiber firings.[ii] The identity between pain and C-fiber firings, like all identities, must hold necessarily if at all.

Kripke now poses a problem. It is at least logically possible, Kripke suggests, to have pain without C-fiber firings, and vice-versa. But if one can exist without the other,

[1] I would like to thank Andy Egan for his invaluable feedback on this paper, including a great deal of discussion, suggestions for related reading on the topic, and comments on drafts.

then the identity between pain and C-fiber firings would be contingent. As we saw above, this identity cannot be contingent because it holds between rigid designators. So there can be no identity between pain and C-fiber firings at all; pain cannot *just be* C-fiber firings.[iii]

This argument does not quite work, for the premise that it is logically possible to have pain without C-fiber firings and vice-versa is a very weak one. I assume that, by 'logically possible,' Kripke means something like, 'consistent with all deductive logical consequences of the given premises.' But what is the relevant premise supposed to be? If the relevant premise is something like, "Pain is feeling *x*," where *x* is picked out by ostension, then even statements like "Pain doesn't hurt" are logically possible. Yet surely it is necessary, in the sense of necessity relevant to the above argument, that pain hurts; something that doesn't hurt isn't pain, in every possible world. This talk of possible worlds points us in the right direction: the sense of possibility that Kripke wants is *metaphysical* rather than logical possibility.

Then Kripke must argue that "Pain is C-fiber firings" is not metaphysically necessary, i.e., that there is some possible world where pain isn't C-fiber firings or where C-fiber firings aren't pains. His argument is the familiar Cartesian one: we can easily imagine pain without C-fiber firings (just imagine a disembodied pain); and we can easily imagine C-fiber firings without pain (consider an otherwise ordinary individual who has no qualia whatsoever, but whose neurophysiology is identical to ours – the familiar 'zombie' case). So Kripke must be committed to some thesis such as the following: *conceivability* implies *possibility*. (Kripke does not have to hold that this works the other way around: it is all right for the argument if possibility does *not* imply conceivability.)

At first blush many of Kripke's own examples appear to be counterexamples to this thesis. Take the necessary identity of water and H_2O. Can't I conceive of a situation in which water is not H_2O? For I can imagine that chemists have known all along that water was composed of sprites, and that they have been conspiratorially hiding this information from the public. One can tell similar stories for the necessary identity of heat with molecular motion. I can conceive that delta-waves rather than molecular motion cause the sensation of heat in me, for instance. It seems that this is a case where heat is not molecular motion, but delta-waves.

Kripke accounts for such examples by retaining the view that conceivability implies possibility, but denying that we are conceiving of what we think we are conceiving of. When I think that I am conceiving of water being composed of sprites rather than H_2O, in fact I am instead conceiving of something *qualitatively identical* to water – something that looks, tastes, feels, and smells just like water – being composed of sprites. This is perfectly possible. What isn't possible is for *water* not to be composed of H_2O. Likewise for heat and molecular motion: it is possible for delta-waves or whatever else to cause the sensation that is in fact caused by heat; what is not possible is for *heat,* for *that very thing*, to be something other than molecular motion. In general, if I think that I can conceive of some situation *S*, either *S* is possible, or I am in fact conceiving of some situation that is merely qualitatively identical with *S* (call it *S′*) and *S′* is possible.[iv]

Maybe the identity theorist can make a move like Kripke's move in the water-H_2O case and the heat-molecular motion case. That is, maybe the identity theorist can say that we cannot conceive of pain without C-fiber firings or vice-versa; rather, we can only conceive of something qualitatively identical to pain without C-fiber firings. *But anything qualitatively identical to pain just is pain*. Pain, at least on one use of the term, is nothing

more than a certain qualitative feeling. So there remains no alternative possibility for the identity theorist to appeal to, no other situation that we are in fact conceiving of when we think we are conceiving of pain without C-fiber firings. Since the Kripkean move fails in the pain case, it is indeed conceivable that there be pain without C-fiber firings; so the identity between pain and C-fiber firings is not necessary; we have already concluded that it cannot be contingent; there must then be no identity between pain and C-fiber firings.

But identity theorists are not the only physicalists on the market. How do other physicalists, like functionalists and eliminative materialists, fare? Functionalists claim that mental states are identical to functional states. Functional states may be roughly defined as some sort of relation among inputs, system states, and outputs; two systems may be composed of radically different substances but still instantiate the same functional state. Unfortunately for the functionalists, Kripke can run an argument that parallels the above argument against identity theory. Suppose the functionalist identifies some type of pain *P* with functional state *F*. To have *P*, the functionalist wants to say, is to be in *F*. There are at least two different ways this claim may be spelled out. *P* might be a type of pain: this sort of pain in my left knee. Or it might just be *pain*, simpliciter. Indeed the functionalist could make both claims and identify the type of pain with functional state F_1, and pain simpliciter with functional state F_2. Of course then every F_1 needs to be an F_2, but not vice versa. Kripke's argument will work just as well against any of these permutations. For we can conceive of pain *P* without functional state *F*, and vice versa. (Just as before, we can conceive of disembodied pain and of zombies.) Conceivability implies possibility; so it is possible to have *P* without *F*; so *P* and *F* are not necessarily identical; so they are not identical at all.

Eliminative materialists do not bother to identify pain with any physical state; they argue that there is no such thing as pain. This is not meant to be the silly assertion that it doesn't hurt when I bang my knee into the table; instead, the suggestion is that the folk-psychological concept of 'pain' has an empty extension. Completed science will furnish a new set of concepts, including no doubt some concept under which will fall the mental state that I am in when I bang my knee against a table. Call this concept *schmain*. Eliminative materialists can attempt to resist Kripke's argument by holding that it is inconceivable that there be schmain without, say, some neurological state. For *being a neurological state* will be part of the concept *schmain*. How well this strategy works is unclear; Kripke will doubtless respond that those things that hurt surely share some (purely qualitative) property. How will the physicalist identify *that* property with anything physicalistically respectable? More can be said on both sides here, but let us suppose that the eliminative materialist can contort her way out of the problem. Then the physicalist can, as a last resort, fall back to eliminative materialism to evade Kripke's argument, but only at the cost of giving up 'pain' as a genuine natural-kind term. This conclusion alone merits rebuttal.

We have so far considered arguments against physicalism. What is alarming is that Kripke's strategy against the physicalist likewise hits many stripes of *dualist*! Consider the 'substance dualist,' who says that mental states are identical to states of some non-material substance. (There are several ways one might explain how a *substance* could be non-material: maybe the substance is non-spatial, or doesn't interact in any way with quarks or whatever the bottom-level physical things are.) Suppose that pain ends up

being identified with M-charges in non-material substance *S*. I can conceive of pain without substance *S*, though (embodied pain), and I can conceive of substance *S* without pain (a soul-zombie)! So it's possible for there to be pain without M-charges in substance *S* and vice-versa … so pain is not identical to M-charges in substance *S*. Likewise for the functional dualist. For, strictly speaking, functionalism need not be a physicalist theory. One might suggest that the relevant functional states are functional states *of non-material substances*. But the above argument against functionalism remains unimpeded.

We might begin to get suspicious here. For Kripke's argument is so strong that it rules out, not only any form of physicalism that takes 'pain' to be a natural-kind term, but also any form of dualism that does the same! The remaining contenders seem to be eliminative materialism and, I suppose, eliminative dualism (dualism of a type that denies the existence of pain) – and even these may not escape Kripke's argument. We might also consider 'anti-reductionism,' the claim that mental states aren't reducible to *anything*, except maybe other mental states. And all of this is supposed to follow from an argument about what we can imagine? Let us keep this suspicion in mind as we turn to the topic of possibility.

II

If we would like to argue that conceivability does not imply possibility in the pain-C-fiber case, a natural first step is to examine the notion of possibility. Perhaps an independent account of possibility will clarify what the conceivability of state of affairs *S* tells us about its possibility.

What is it for state of affairs *S* to be possible? I take it that the following paraphrase is uncontentious: if *S* is possible, then there is a *way things could be* such that *S* obtains. Replacing the italicized phrase with another: if *S* is possible, there exists *an arrangement of things* (which need not be actualized) such that *S*. We might turn this intuition into a formalized theory as follows. Call complete possible situations *possible worlds*. Then let possible worlds be composed of things arranged in certain ways relative to each other.

A quick aside on the ontological status of possible worlds: one might think, with David Lewis, that possible worlds are things that are really out there, with various events really going on in them, in the same sense that our world is really here and events are really going in our world. Alternatively, one might claim that possible worlds are abstract objects of some kind, or sets of sentences/statements/propositions, or both. I prefer the latter approach for reasons of ontological parsimony, and I will accordingly tell a story about possibility using the ontologically deflationary variety of possible worlds. But my argument goes through just the same if one takes Lewis's view of the ontological status of possible worlds.

Here is one way of explaining what goes into a possible world. Let possible worlds be abstract objects consisting of mathematical representations of space and time. Consider an *n-dimensional grid*, which represents spacetime. It has whatever features our completed physics will attribute to spacetime: perhaps n-1 dimensions are spatial and the remaining dimension is temporal. Consider also a set of *micro-level entities*. Let these be whatever the bottom-level particles of the actual world are: quarks, gluons, strings, or whatever is in fact the particle or particles from which all other particles are built. A

possible world is just an n-dimensional grid populated with these micro-level entities. We can construct all possible worlds using the following rules:

1. The actual world is possible.
2. If world w_1 is possible, then so is a world w_2 which is exactly like w_1, except that exactly one of the following obtains:
    a. Any single empty spacetime-point in w_1 instead contains any single micro-particle in w_2.
    b. Any single non-empty spacetime point in w_1 is empty in w_2.
    c. Any single spacetime point in w_1 does not exist in w_2.

At least one more rule is needed to add spacetime points. Exactly how this rule is phrased depends on facts about the geometry of spacetime about which I plead ignorance. But I doubt that, given the mathematical formulation of spacetime in completed physics, it will be particularly difficult to develop such a rule.

There are many important objections which can be raised against such a theory, most of which are outside the scope of this paper. Let me mention just a few objections which are especially relevant for my purposes here. First, one might have a worry about certain terms I have used above. Suppose for a moment that final physics tells us that there is one fundamental particle in the universe: strings. It follows from my account that there are possible worlds with things I call 'strings' which, the objection might maintain, are not strings at all. For suppose that, in the actual world, strings follow some set of laws *L*. Kripke points out that natural-kind terms, like 'string,' are rigid designators, so they pick out the same thing in every possible world; and presumably, it is essential to strings that they obey the set of laws *L*. But in some possible worlds, things that I call 'strings' fail to obey these laws (e.g., perhaps a string spontaneously disappears at some time *t* when *L* would not permit such a disappearance).

Point well taken. I need to have some replacement for the various terms I use to pick out micro-particles: I can't say 'string,' 'quark,' or whatever. Instead, let me use some suitable mathematical description for each micro-particle, and replace every instance of the name of the relevant micro-particle with this description. Obviously, the mathematical description will have to omit any claims about how the micro-particle behaves over time; it will confine itself to a description of how large the particle is, whether the particle has (let us say) positive or negative charge, etc. (Or rather, positive or negative 'charge,' lest the objection recur for my use of that term. Again, some kind of time-independent mathematical description will have to be given for each property that characterizes the micro-particle.)

Next objection. One might think that this account is too restrictive to capture all the metaphysically necessary truths. For couldn't spacetime have had a different number of dimensions? And couldn't there have been different fundamental micro-particles? In response to such worries, one might broaden the generation rules for possible worlds. Suppose we have our theory from completed physics. It tells us that there is a world of n-dimensional spacetime with some number of micro-particles; when we ask for details about the natures of these micro-particles, the physicists give us some mathematically formalized properties. I contend that there will be natural ways of extending these mathematical formalizations so as to define new particles, none of which actually exist.

Likewise for spacetime: we can develop natural extensions of the mathematical formalization which describes spacetime. We could then rewrite the possible world generation rules so as to capture these extensions. This project is not feasible now for the simple reason that we don't have completed physics before us.

One final objection with direct relevance for my thesis of the possibility of physicalism. Doesn't my account just *beg the question* against anti-reductionists? The emergent dualist, for instance, will say that two worlds might have exactly the same micro-particles in exactly the same places throughout spacetime, and yet differ in their mental properties. I concede immediately that the emergent dualist will have to subscribe to a different view of possibility. (Though she can go in for something that looks a lot like my view; she will have to modify 2a-2c by, say, adding qualia-generation and qualia-elimination rules.) But my goal in this paper is merely the modest goal of showing how one might plausibly meet the Kripkean objection. It is not my concern to *refute* emergent dualism, but only to show how an identity theorist, functionalist, etc., might defend himself via a certain view of possibility.

III

I hope we now have a better handle on what it is for a situation to be possible: there must be a possible world of micro-particles arranged in spacetime such that the situation obtains. Given the above account of possibility, what does the conceivability of a situation tell us about its possibility? Before we can answer this question, we have to know what *conceivability* is. Let me avoid a fight and grant my opponent the most charitable reading: let us take conceivability to be *vivid imaginability*. This reading of conceivability is a particularly strong one; and I will also grant that we can conceive, in this strong sense, that pain exists without C-fiber firings and C-fiber firings without pain.

Conceivability is good evidence for possibility exactly when our ability to conceive of a situation *S* is good evidence for there being some possible arrangement of micro-particles such that *S* obtains. So here is a case in which conceivability seems to provide very strong evidence for possibility. I can conceive of (vividly imagine there being) things like horses which have two antlers emerging from their heads. Call these creatures *bicorns*. The fact that I can conceive of bicorns makes it very likely that it is possible for there to be bicorns. For, having seen horses before, I know that horses are possible – that there exists *some* arrangement of micro-particles, I know not which one, that forms a horse. And, having seen antlers before, I know that they are also possible, for there is at least one arrangement of micro-particles that makes up an antler. So I know that it is possible for there to be bicorns, since all that requires is having the antler-micro-particle arrangement placed twice, in the appropriate places, atop the horse-micro-particle arrangement. Moreover, my vividly imagining a bicorn is what reveals to me that bicorns are possible: it makes perspicuous the spatial relationships that need to obtain between the antlers and the horse. Put another way: it tells me how to construct the possible world in which there are bicorns. I begin with the actual world and pick out some horse, and then I twice replace some of the micro-particles above its head with the antler-micro-particle arrangement, using rules *a* and *b* of the possible-world generation rules.

Inconceivability may also be good evidence for impossibility. I cannot conceive of anything – say, the lamp on my desk – that is both red all over and green all over.

When I try, I realize that the closest I can come is imagining the lamp red all over, and then imagining it being green all over. Now there is some arrangement of micro-particles *R* that forms a red-all-over lamp, and there is a different arrangement of micro-particles *G* that forms a green-all-over lamp. But there cannot be both at the same time in the same place, and the possible-world generation rules tell us why not. In the actual world, I assume, there is no time and place at which two particles exist; and the possible-world generation rules never replace one particle in a spacetime point with two particles. Again, the fact that I cannot conceive of something being both red all over and green all over is what gives me good evidence for the impossibility of such a situation obtaining, since the reason that I fail to be able to imagine the situation vividly is that I cannot imagine two things in the same place at the same time.

But inconceivability is not always good evidence for impossibility. One example: I cannot conceive of things being the way they actually are. Given my very limited cognitive capacity, I can't vividly imagine everything in the entire universe, even if I had before me a complete description of it. But my failure to be able to conceive of things being how they actually are doesn't yield any evidence that there is no possible arrangement of micro-particles such that things are as they actually are. Even though I can't conceive of how everything actually is, I can still easily tell you how to generate all of actuality from the possible-world generation rules: stop at step 1.

Recall Kripke's account of modal error: when we conceive of situation *S*, either situation *S* is possible *or some qualitatively identical situation is possible*. The italicized bit explains how we mistakenly think we can conceive of water that isn't H_2O, or heat that isn't molecular motion. I offer a different account of modal error. For (alleged) situation *S* to be possible is for there to be some arrangement of micro-particles such that *S*. There are two ways that we might mistakenly take *S* to be possible. First, there might be some possible situation *S′* which we mistake for the (impossible) situation *S*. Second, there might be no arrangement of micro-particles such that *S*. The first type of mistake is my analysis for the water-without-H_2O case and the heat-without-molecular-motion case. The second type of mistake is my analysis of the pain-without-C-fiber-firings case. Let us take these one at a time.

There are possible situations (arrangements of micro-particles) where there is stuff that is superficially like water, but isn't H_2O. There are also possible situations (arrangements of micro-particles) where there is something that feels like heat, but isn't molecular motion. It is a fact about language – about what *counts as* water – that the superficially-water-like stuff isn't water, and it is also a fact about language that the thing that feels hot, but isn't molecular motion, isn't heat. We conceive of possible situations with water-like and heat-like things and incorrectly take them to be situations where water and heat fail to have their actual deep structures. The mistake is understandable, as it merely involves confusing *things recognized in the same way that we recognize the thing picked out by a term* for *the thing that is actually picked out by the term*.

Now for the second type of mistake, and, at last, the rebuttal of Kripke's argument. I need to show that the fact that I can imagine pain without C-fiber firings, or functional state F, or whatever, lends no support to the claim that there is a possible arrangement of micro-particles such that there is pain without the relevant physical state. What happens when I imagine pain without C-fiber firings? I imagine a person with no C-fibers in her head – just silicon, say – and then work myself into a state of mind where

I'm imagining pain. I then say to myself, "All right, *that* sensation is what she is feeling." But that doesn't tell me how to construct, using the possible-world generation rules, a situation where there's pain but no C-fiber firings. At best it tells me how to construct a situation where there's a person with silicon in her head; it leaves me baffled about what micro-particles to add, and where to add them, for there to be pain, too. I may be mislead by my vivid imaginings into believing that there is some arrangement of micro-particles such that pain-without-C-fiber-firings obtains, but *there needn't be any such situation*.

Likewise for conceiving of C-fiber firings without pain. I can imagine myself in a brain state where my C-fibers are firing, and I can add to that picture the thought that I feel no pain. But that doesn't tell me how to construct a situation where my C-fibers are firing without my feeling pain; the best I can do is reproduce the micro-particle state that I am actually in when my C-fibers are firing, and by hypothesis whenever that *actually* happens I'm in pain! Again, the physicalist can claim that there really is no situation where there's C-fiber-firings without pain, and the conceivability argument does nothing to show otherwise.

This tack might be taken as question-begging. I have packed into my notion of possibility that there can be no difference in the world without a difference in micro-particle arrangement, and this could be taken as already begging the question against the Kripkean. So the dedicated Kripkean need not be persuaded by my argument here. Even if that is the case, my argument does something important in response to Kripke's objection: it shows that there is a coherent, and indeed very attractive, alternative. My alternative (1) preserves our intuition that conceivability is a good guide to possibility in many important cases (and inconceivability a good guide to impossibility), (2) gives us a way of telling when conceivability isn't a good guide to possibility (and similarly, *mutatis mutandis*, for inconceivability and impossibility), and (3) lets us be physicalists and allows science, rather than *a priori* philosophy, to resolve the apparently empirical question of what pain and other mental states are. The force of Kripke's objection lay, in part, in the apparent difficulty of providing any satisfactory alternative account of the relationship between conceivability and possibility. I claim to have provided one.

---

[i] Saul Kripke, *Naming and Necessity* (Harvard University Press: Cambridge, 1972), p. 3.
[ii] Ibid, pp. 148-9.
[iii] Ibid, pp. 146-9.
[iv] Ibid, pp. 99-104.