Lewis's Epicycles, Possible Worlds, and the Mysteries of Modality*

Elijah Millgram

Department of Philosophy
University of Utah
Salt Lake City UT 84112
lije@philosophy.utah.edu

July 12, 2011

If there's an obvious lesson to draw from twentieth-century metaphysics, it's that reductionism doesn't work. Reductions purport to show that things of one kind (minds, for instance, or physical objects) are *really just* things of another kind (dispositions to behave in certain ways, or patterns in the flux of sensation, respectively), the proof of the pudding being that you could give a scheme for paraphrasing away anything you might say about the first kind of thing into assertions about the second kind of thing. Over the course of the century, reduction after proposed reduction failed, and failed systematically, that is, for the same small family of reasons. So it's

*I'm grateful to Irene Appelbaum, Aisling Crean, Kate Elgin, Don Garrett, Kathrin Koslicki, Beatrice Longuenesse, Cei Maslen, Fraser McBride, Shaun Nichols, Doug Patterson, Laurie Paul, Guy Rohrbaugh, Candace Vogler and Matt Weiner for helpful conversation, to Chrisoula Andreou, Sarah Buss, Ben Crowe, Christoph Fehige, Alan Hajek, Clif McIntosh, Laura Schroeter, Scott Shalkowski and Mariam Thalos for comments on earlier drafts, and, for valuable feedback, to audiences at Brigham Young University, Kings College London, the University of Reading, the University of Colorado, the University of Notre Dame and the University of the Saarland. Thanks also to the University of Utah's College of Humanities for sabbatical support.

something of a surprise that the position we find serving as perhaps the most visible reference point within turn-of-the-millenium metaphysics is yet another form of reductionism: this time, from *modal* facts to configurations of 'possible worlds'.

David Lewis called his view "modal realism," and as I just remarked, it has come to serve as something of a reference point. But it is also seen as an extreme and idiosyncratic position, and my real concern, for which I mean to use Lewis's view as a stalking horse, is the much more widespread idea embedded in it, that statements with modal content can be paraphrased into the vocabulary of possible worlds. For this reason, and partly as well because Lewis was such an elaborate system builder, I am going to divide up my discussion of his modal reductionism into two treatments. The other one takes up his realism proper, that is, the insistence that possible worlds, global ways things might have been, exist (*really, really* exist, in just the way anything else does); it addresses itself to a specifically metaphysical motivation, namely, the conviction that modal facts are, on the face of it, spooky, unnatural, and have no proper place in the physical world.[1] Here I will focus on a different but equally important motivation for reduction, the worry that you do not understand the very words you are using. (When reductions are motivated in this way, they tend to get thought of as conceptual analyses.) The divide-and-conquer approach means that I need to ask for a slightly unusual concession from my readers: If you think of an objection to the argument, and you do not see it addressed here, please check the companion treatment before deciding that I have overlooked one or another of the resources provided by Lewis's metaphysics.

I'll begin by introducing the notion of modality, and by describing Lewis's modal reduction. I will take time out to explain what is required of a reduction, to document that Lewis accepted the requirements, and that he has been widely read as doing so. At that point I'll lay out one of the arguments that buried some earlier twentieth-century reductionisms, and construct a variant directed against Lewis's.

Lewis's reductionism has been criticized on occasion, and on the way in I will distinguish the argument I am using to kick off the discussion from those more familiar complaints. Once again, my intent is to use Lewis as a stalking horse, and to show a widely shared assumption to be false: that ordinary claims with modal content can be paraphrased in the vocabulary of possible worlds. Because the extant criticisms share this assumption, they do not do the work that we need.

[1] Millgram, 2009, ch. 11.

Although Lewis is in certain respects an unusual representative of the take on modality I want to contest, his position is perhaps the most thoroughly articulated and inventively worked out in the neighborhood. So I will consider at some length whether his systematic metaphysics supplies the materials for a rejoinder to my argument. As it will turn out, it does not, but, more importantly, this stretch of counterargument will put in place two secondary conclusions, one regarding the provenance of the materials upon which we draw when we assess counterfactuals, and the other having to do with the fragmentary nature of those materials. If these conclusions are correct, they are important guides for investigations of modality down the road.

Recall that we will have been focusing on a reduction that is motivated as conceptual analysis, and leaving what philosophers nowadays tend to insist are the properly metaphysical considerations for elsewhere. It is nonetheless important to see how these apparently different modes of treatment fit together, and in sec. 10 I will take up the connections between these two philosophical motivations. I will rehearse the reasons that modal subject matter does not allow a graceful retreat to the fallback position eventually adopted by so many other twentieth century reductionist programs: that supervenience will give us everything we wanted from a reduction, but couldn't get. I will explain why the missing fallback preempts what has elsewhere become a standard move, namely, to insist that what is on offer in one's theory is not meant as conceptual analysis, but is rather nonreductionist metaphysics or ontology.

I'll wrap up by asking what the big-picture lesson of the argument is. Discussions of modality have come to take the possible-worlds way of thinking about it for granted. The philosophical uses of the possible-worlds apparatus depend on the assumption I intend to refute, that we can give possible-worlds renderings of ordinary claims with modal content. I will conclude that it is time for a dramatic shift of key in philosophical thinking about modality.

1

Like many metaphysically deep phenomena, modality is hard to define in a noncircular and illuminating way, and for now I'll introduce the notion with a gesture rather than a definition. Think about claims like these:

1. I *could* have been a contender.

2. Ducks don't smoke cigars, but they *might've.*

3. If Clinton *had* had a decent haircut, he *would* never have been elected Governor of Arkansas.

4. Water *has* to be H_2O.

5. If you *were* King Mithridates, you *would have* foiled your enemies' evil plans.[2]

What they have in common is modal subject matter (marked, in these sentences, by the highlighted *could*, *would*, *has to*, *might've*, and so on). To a first approximation, the subject matter includes the possible and the necessary, counterfactuals (represented by sentences of the form, 'If such-and-such *had* happened, so-and-so *would've* happened'), dispositions (because they normally embed counterfactuals: if the glass is fragile, then if you had dropped it, it would have broken), and perhaps other things of the same ilk (whatever ilk that is) besides.

Modality is philosophically mysterious. When you assert that ducks smoke cigars, I know just what to look out for to confirm or disconfirm your claim. The sentence lists the relevant items (ducks, cigars and smoking), and all of these are things that I'll recognize when I see them. But when you tell me that ducks *might've* smoked cigars, there's a part of the sentence—the '*might have*'—that leaves me wondering what to look for, and how I would recognize it if I found it. Consequently, it's easy to start worrying that I don't know how to say what would make such a sentence true (except trivially, by saying that it would be true if ducks indeed might've smoked cigars), and if you think of being able to give a sentence's truth conditions as a touchstone for whether I know what it means, then it looks like I don't actually understand sentences that contain *mights*, *would've*s, and so on.

Lewis's modal reductionism was meant as a philosophical response to the mystery, and it was a natural move to make given the historical background. Modal logics—formal calculi intended to represent possibility and necessity—had gained in popularity and prestige after model-theoretic treatments of a handful of the axiomatic systems seemed to put them on a mathematical footing with more traditional Frege-Russell logic. In these so-called possible-worlds semantics, each possible world (intuitively, each way things might have been) was represented by the sort of model of the universe of discourse that would have been used to model a theory in standard first-order logic. "Possibility" was then rendered as a sentence's being satisfied at some world-model, and "necessity" as the sentence's being satisfied at all world-

[2]Props to Kazan, 1954, Gerber *et al.*, 2008, Putnam, 1975b, Housman, 1965, p. 90.

models.[3] Once you do that, "Ducks might've smoked cigars" gets paraphrased as, "There are some possible worlds in which ducks smoke cigars." The "might've" has been replaced by the unmysterious "there are some," plus, of course, the possible worlds—and although these might sound mysterious themselves, we do ordinarily talk about different ways things might have been.

There's nothing to get philosophers to jump on a bandwagon like a formal representation with mathematical cachet, but there was more to it than that: model-theoretic semantics for modal logics made logicians comfortable with modality, in very much the same way that Cantor's work had made mathematicians (and philosophers) comfortable with infinity; where before the subject matter had seemed rife with contradictions and incoherences, now one had an internally consistent way of talking about it, and one that gave you an almost mechanical way of answering questions about (for instance, and don't worry if you don't know what this means) iterated modal operators.[4] The philosophers quickly came to accept "true in all possible worlds" and "true in some possible worlds" as explications or paraphrases of "necessary" and "possible", respectively, and Lewis extended the way of speaking to counterfactual conditionals (again, sentences like, "If so-and-so were the case, then such-and-such would be the case"): if the possible worlds are embedded in a similarity space (that is, if we say that worlds are 'nearer' to one another if they are more similar), then the counterfactual conditional is true if, roughly, in nearby (or the nearest) worlds in which its antecedent is true, its consequent is true. So to determine the truth of a sentence like, "If hybrids were cheaper, more people would buy them," notionally travel out to the 'nearest', i.e., most similar, possible worlds in which hybrids are cheaper, and if, in those worlds, they get bought by more people, the counterfactual is true.[5]

[3]More carefully, at *accessible* world-models; the semantics included a specification of which worlds are, again intuitively, 'visible' from which other worlds; different specifications of the accessibility relation model different axiomatic systems. Formal work on modal logics began with an attempt by C. I. Lewis to represent the non-truth-functional conditionals he needed for his phenomenalist reductionism, about which more below. The model-theoretic approach was developed in work by Saul Kripke and others; for textbook surveys, see Hughes and Cresswell, 1996, Forbes, 1985, chs. 1–2, Priest, 2001, chs. 2–4.

[4]We should always remember that having a representation of the claim that p (even an elegant mathematical representation) does not count as a philosophical argument for p. (Not *at all*: merely to *say* that p, no matter *how* you say it, is not an argument for p.) However, I'm grateful to Tim Bays and Alasdair MacIntyre for pressing me not to forget the reasons that possible-worlds semantics became so popular in the first place.

[5]See Lewis, 1986, Lewis, 1973. That 'roughly' is there to sidestep complexities introduced when the ordering of possibilities by 'nearness' does not have maximal elements;

Then Lewis took the step that turned him into the Alexius Meinong of the twentieth century, which was to treat the possible worlds as *things out there*.[6] This struck many other philosophers as outlandish, since possible worlds are, Lewis acknowledged, spatio-temporally isolated from each other—meaning, no matter how far you travel, or how long you wait, you'll never get to another way things might have been, and so you can't ever *inspect* one of these entities. But Lewis took the move to be justified by its philosophical payoffs: once you have these entities, you can treat the paraphrase of 'necessary' as 'true in all possible worlds' as a *reduction*: not just useful idiom borrowed from the mathematical logicians, but a description of what's really going on when you (correctly) say that something had to happen. You can likewise explicate 'actual' as an indexical, one which picks out the world the speaker is in, in something like the way that 'I' picks out the person who says it; likewise for 'possible' and 'true in some world'. And, very importantly, likewise for Lewis's proprietary semantics for counterfactuals: what it *is* (*all* it is) for it to be true that if Clinton had had a decent haircut, the folks in Arkansas would never have elected him, is that, in the nearest possible worlds in which he had a decent haircut, they didn't. The mysteriousness of modality is addressed: the 'would' in 'they wouldn't have elected him' picks out an object, or, in this case, a class of objects, in just the way that 'Clinton' and 'haircut' do: namely, the relevant class of possible worlds.

Lewis's view is straightforward-sounding, but it has its complications, and because one of them will be important shortly, let me describe it now. Take a counterfactual such as, "I could have been a contender." On the way of talking that Lewis made into a reduction-schema, that gets paraphrased as, "In some (at least one) possible world, I am a contender." But does that mean that *I* am an inhabitant of this other world—that I am both here and in some other place, one that's *so* very far away that it doesn't even count as far away?[7] Lewis handled this difficulty by adopting what he called 'counterpart theory': the contender in that other possible world is

see Lewis, 1983–1986, vol. 2, pp. 6–10.

[6]Lewis took the trouble to distinguish his own doctrines from Meinong's (1986, pp. 98f), something that would not have been necessary if the resemblance had not been hard to miss. The differences he pointed out do not defuse the characterization: to be the Meinong of a given philosophical period is not the same thing as having precisely Meinong's views.

[7]Specialists will be aware that that way of putting it is a detour around a rather different-sounding argument (Lewis, 1986, pp. 198ff), and the back-and-forth that it generated in the literature. I am taking the expository shortcut because, in my judgment, the moves on both sides were badly motivated. For further discussion, see Lewis, 1973, pp. 38f.

not, exactly, *you*, but someone (your counterpart) who is relevantly similar to you, both intrinsically and terms of the role or position he or she occupies in that world. So a counterfactual like this one ends up taking on a further layer of paraphrase: "In some possible world, I have a counterpart who is a contender."

So much for our brutally rapid introduction to Lewis's metaphysics of modality.[8]

2

If your philosophical motivation for a reduction is the worry that you don't understand your own vocabulary, in this case, the *coulds*, *woulds*, *musts* and so on, surely that worry is only assuaged if the paraphrase you provide doesn't contain the vocabulary you suspect you don't understand: if I do not understand the "outgrabe" in Lewis Carroll's "Jabberwocky," it does not help to tell me that it's what mome raths do, because I do not understand "mome raths" either. (In fact, there's an alternative, which is, more or less, to exhibit the relations between the terms in the problematic vocabulary, and how the group of interrelated terms is collectively related to unproblematic objects or notions; this approach has come to be called the Canberra Plan, and I'll take it up in due course.) A reduction of one sort of vocabulary to another sort of vocabulary commits itself to specifying how statements apparently about the former can be paraphrased without residuum into statements about the latter, and without importing, explicitly

[8]Not only I am leaving the full discussion of Lewis's realism to another treatment, I am also going to put to one side a further and very obvious problem. Earlier on, I excused myself from giving a proper definition of modality, and Lewis himself, in the book-length presentation of his view, said almost nothing about what modality is (even though one of the advertised benefits of the view is an analysis of modality). That's very peculiar, because if you were going to give an argument that As are really Bs, you'd think that such an argument would have to turn on an independent characterization of the As; without such a characterization, Lewis *could* not have a good argument for his own central claim: that modal facts are really just facts about the possible worlds.

For instance, one view people often have is that 'metaphysical modality' is a different sort of thing from 'epistemic modality', and an account of the former needn't include an account of the latter. (In sentences in which, for instance, "might be" and its relatives mark epistemic modality, they are paraphrasable, with some qualifications at the margin, by variants of "for all you know". So, "He might be home by now" would get rendered as, "For all I know, he's home by now.") Lewis seems to think otherwise, because he promises a treatment of the epistemic modals as well (1986, pp. 27ff); but without an explanation of what you meant by 'modality' in the first place, how could you know whether such a treatment was owed?

or surreptitiously, the suspect vocabulary into the paraphrase. That does not disallow using the concepts and vocabulary one is trying to replace in the course of *identifying* the paraphrase, but those concepts and vocabulary had better not end up in the reductive paraphrase itself. A reduction also commits itself to eliminating black boxes, by which I mean, devices in the conceptual analysis whose inner workings it cannot exhibit.[9]

When we get to sec. 10, I will, as promised, lay out the play of forces that committed Lewis to a reduction. But reductionism is so unfashionable nowadays that even compelling reasons may fail to dispell the worry that I am being interpretively uncharitable. For that reason, before we go any further, I want to document both that Lewis himself accepted the demand, and that his interlocutors have read him as accepting it. (If this isn't something that's bothering you, you're welcome to skip ahead to the next section.)

As I mentioned earlier on, Lewis's position has already attracted criticism, and the relevant complaints turn on two related worries. One was that Lewis had no explanation of what made his other possible worlds *possible*: since they were, on his account, just like this world, weren't they merely more *actual* worlds?[10] The point of the objection is that an unexplained grasp of possibility must in fact be concealed in the appeal to possible worlds, that if our concern is that we do not understand modal-

[9]Here is a toy example of a black box: "A sentence containing modal vocabulary really means whatever David Lewis says it means." Even though that last phrase contains no *woulds* or *coulds*, unless I know what Lewis's gloss is, I'm not in a position to tell whether it really is a content-preserving paraphrase, and I'm not in a position to tell whether or not it reintroduces the dicey terms and concepts we are concerned we do not control. (For all I know, what Lewis says it means is full of 'coulds' and 'woulds'.) That was, again, a toy example, but appeals to black boxes play a large and disreputable role in much recent philosophy, as when moral philosophers invoke the preferences agents would have in idealized circumstances; for discussion, see Millgram, 2005, pp. 69, 74, 85n39. To foreshadow, secs. 5–8 will be devoted to a black box in Lewis's treatment of modality, namely, similarity relations among possible worlds. Another of the black boxes—the context-dependence of such similarity orderings, thus also of the counterpart relation—will come in for discussion in the notes.

[10]Complaints in this ballpark include Shalkowski, 1994, Plantiga, 1987, Sider, 2003, sec. 3, and Chihara, 1998, pp. 280, 286 (see also p. 80, which fields a similar complaint regarding Lewis's 'worldmate' relation). Divers, 1997, discusses the interplay between the two classes of objections; Divers, 2002, ch. 7, provides an overview of the back and forth around the former.

There have been other attempts to show attempted reductions of modality to rely on undischarged modal notions. For instance, MacBride, 2001, primarily discusses Jubien, and focuses on the modal content of metaphysical categories such as "property," "object," and "matter": an object *cannot* be instantiated, matter *must* be spatially and temporally located, etc.

ity, we should be equally concerned that we do not know what a possible world is, and that the obligations of the reductive paraphrase have not been met. That is, these objections take it for granted that Lewis is to be read as attempting a reduction. (Bear in mind that our present concern is not whether the criticisms are effective, but only how they were motivated.)[11]

When Lewis addressed this complaint, he nicely exhibited the reductionist shape of the project. Lewis reiterated that his indexical understanding of 'actual', on which it picks out *this* world, entailed that other worlds could not be actual, and it must have seemed to his opponents that he was making a point of missing the point of their objections.[12] But his response exhibited his reductionist commitments. Since the object of the exercise was to reduce away the modal notions, those notions should not appear in a characterization of the items to which they were being reduced: for unreduced mere possibility to reappear as a feature of those other worlds would mean that the reduction had failed.[13]

The second family of objections had it that which possible worlds (or occupants of possible worlds) there are determines what comes out true when you quantify over them; but how can you say which there are other than: the *possible* ones?[14] As before, the force of these objections was that the possible-worlds rendering surreptitiously deploys unreduced modal concepts and primitively modal opinions, and so it does not reduce them. So it again exhibits the agreement of Lewis's interlocutors that he was to

[11]Indeed, there are discussions that more or less explicitly describe Lewis's project as I do. Sider, 2003, and Plantiga, 1987, have also called Lewis a reductionist—though in Plantiga's case, the label is attached to a somewhat different point. Divers, 1997, p. 144, is willing to use the word. Sider, Divers, 2002, p. 106, and Chihara, 1998, pp. 81f, 207n, agree with me on the substantive claim—though Chihara does not use the term himself.

[12]Lewis, 1986, pp. 97ff. Notice that Lewis's response predates the publications listed in note 10, which makes the belated and reiterated complaints an indication of just how hard it was to swallow.

[13]Here's some further textual confirmation of the stance. Lewis treated it as an objection to competing views that they had to help themselves to 'primitive modality' (Lewis, 1986, pp. 151–55, 167f, 179f), and he stated flat out that "[p]rimitive modality is bad news" (p. 242). Compare Lewis, 1999, p. 298.

[14]See, for instance, Melia, 2003, pp. 114f, Lycan and Shapiro, 1986, p. 358, Divers and Melia, 2002 (where the claim focuses on alien properties), Bremer, 2003 (a rebuttal treating individuals the way Lewis had wanted to treat the worlds), Divers and Melia, 2003 (refielding the initial complaint, only now about possible individuals). Chihara, 1998, pp. 282f, considers the worry that a reductionist analysis of modality, constructed with an eye to making one's modal views come out right, is not reductionist *enough*. Paseau, 2006, p. 724, briefly entertains the reply (which he does not endorse) that the range of possible worlds should be taken to be the right one, whether we can specify it or not. Divers, 2002, ch. 7, is a recap of previous exchanges.

be read as committed to a reduction of modality.

And as before, Lewis confirmed that commitment in his own counterarguments; he responded by attempting to specify the range of possibilities as ways of recombining elements of the actual world.[15] As before, our interest is not in whether his rejoinder was successful, but in what it shows about what he was trying to do. By attempting to address the complaint on its own terms, he was accepting those terms: because the range of possibilities operated in his account as a black box, he needed to show that it could be opened up, and its workings reconstructed using only modality free materials.[16]

As announced, I intend to sidestep this back and forth, because it takes for granted the assumption I mean to contest, and which I am using Lewis's view to reconsider. You can believe that understanding what a possible world is requires a primitive grasp of possibility, and that quantifying over possibilia gives you real results only if you have prior opinions about what is possible, while *still* agreeing that the content of a possible-worlds paraphrase is that of the modally loaded claim it purports to render. Instead, I propose to consider counterfactuals. These are the hardest-working division of our modal apparatus; ordinary people do not have much in the way of strong opinions about what is possible or necessary, but they have a great many opinions, opinions on which they rely in their day-to-day lives, about

[15] Lewis, 1986, pp. 86–90.

[16] This is as good a place as any to speculate about why Lewis called his view 'modal realism', not (as I've been characterizing it) 'modal reductionism'. Even if realism and reductionism are opposites of a sort, you can be both realist and reductionist if you are a realist about one kind of thing, and a reductionist about another; normally, as Putnam once remarked, reductionist projects presume that whatever they are reducing their problematic items to *are* real. (Thus, a phenomenalist is normally a realist about sense data.) Reductionist projects are often epistemologically motivated, and when they are, they follow the epistemic order: that is, what you're reducing *to* is whatever is easier to know. (If you think that material objects are to be reduced to sensations, that's because you take sensations to be the sort of thing of which one is immediately aware, and so because you take yourself to be solving the skeptical problem of how one ever knows anything about material objects.) When reductions are not epistemologically motivated, they're often driven by a prior picture of what exists on its own, or in its own right. (If you think that biological systems are to be reduced to physical systems, that's because you take physical objects to exist in their own right, but biological objects to exist only by grace of the configuration of physical objects.) Lewis's modal reductionism follows neither the epistemic order (again, a typical complaint about the view was that you can't observe other possible worlds) nor the usual views about what exists all on its own (it's not standard to think of the might-have-beens as the substances). So Lewis may not have found it natural to describe himself as a reductionist because the starting and ending points of the reduction he was proposing were in these respects unusual.

what would have happened, if... Accordingly, the objective of the coming argument will be to demonstrate that counterfactuals cannot be paraphrased into the vocabulary of possible worlds. If the argument is successful, it will show that Lewis's reduction fails, but, and more importantly, it will do so by showing that an especially important class of modally laden claims cannot be given a possible-worlds paraphrase.

If the worry you are trying to defuse is that you don't understand those *coulds*, *woulds* and so on, you have to show that (perhaps with some adjustments at the margins) you do understand the very *coulds*, *woulds* and so on that people actually utter. And this is (one reason) why the reduction Lewis was considering was a modality-free paraphrase of, specifically, ordinary modally laden claims about ordinary things.

3

Lewis used to complain that the most frequent response to his view was an incredulous stare, but that an incredulous stare isn't an argument.[17] We *will* get an argument; before we start in on it, however, here's a bit of warmup.

Early on in the back-and-forth of twentieth-century metaphysics and epistemology, Roderick Chisholm laid out the objection that I mean to use as a pattern, the occasion being reductionist formulations of phenomenalism. On phenomenalist views such as those of C. I. Lewis,[18] ordinary material objects are really nothing over and above patterns in the flux of actual and possible sensation, and the content of a statement about such an object, for instance, "The cat is asleep on her pillow," is given by many, many counterfactuals along the lines of, "If I were to have a sensation like *so* [the one I normally describe as the feeling of turning my head], I would come to have a visual sensation like *that* [the one I would normally describe by saying that it's of a cat sleeping on a pillow]"—that is, by counterfactuals about the sensations.

Leave to one side obvious difficulties about stating these counterfactuals without mentioning material objects (pillow, cat,...) in the course of picking out the sensations; waive the objection (since C. I. Lewis himself happily

[17] The characterization made it into a widely-circulated collection of humorous 'proofs that p' (compiled by Hartry Field); see Lewis, 1986, p. 133, and Lewis, 1973, p. 86, for the complaint.

[18] Clarence Irving, not to be confused with David Lewis; I'll always give the earlier Lewis's initials, and a freestanding "Lewis" will always be David Lewis. For the phenomenalism, see C. I. Lewis, 1956, and C. I. Lewis, 1946; for the objection, see Chisholm, 1948.

conceded it) that there are too many such counterfactuals to list—perhaps infinitely many. Chisholm focused instead on the hard-to-miss fact that pretty much any such counterfactual is only true other things being equal (or *ceteris paribus*, in philosophers' Latin). For instance, while it's true, other things being equal, that if I were to turn my head (or have that head-turning sensation), I would see (or seem to see) the sleeping cat, nonetheless, if someone were to smash me over the head with a sledgehammer, just as I was starting to turn, then I *wouldn't* see the cat. Now we might try to accommodate such facts by adding extra clauses to the counterfactuals: "If I were to have a sensation like *so*—and no one were to smash me over the head with a sledgehammer—I would come to have a visual sensation like *that*." But, and this is the problem, the *ceteris paribus* clauses are stated in the vocabulary of material objects (this one mentions a sledgehammer), viz., in the very vocabulary that the reductionist phenomenalist is committed to paraphrasing away.

To be sure, you could try rendering the bit about the sledgehammer into further counterfactuals about sensations, but these will require their own *ceteris paribus* clauses, once again couched in material-object vocabulary: If I were to have the picking-up-the-sledgehammer sensation, I would feel the heft... unless a nerve were cut... and so on, down the regress. The reductionist promise of paraphrasing away the vocabulary of material objects turns out to be empty.

Here's one more quick dress rehearsal, another argument on the same pattern due to Hilary Putnam. Logical behaviorism is the view that statements framed in psychological vocabulary can be translated into (or at least analytically entail) statements framed in a non-psychological behavior-description vocabulary; for instance, if you're in extreme pain, then you're more likely to scream. Now behavior is produced by indefinitely many psychological states jointly, and in particular, sometimes we suppress behavior prompted by one psychological state on the basis of another. Putnam memorably called the characters in his illustration 'super-spartans': extreme pain makes super-spartans want to scream, just like anybody else; however, the super-spartans want, even more strongly, not to show the pain, and so they hold it in. The relevant entailments or translations only hold if other things are equal, and there are always going to be *ceteris paribus* clauses that have to be stated in psychological vocabulary. As before, the reductionist promise of logical behaviorism turns out to be empty.[19]

[19]Putnam, 1975a; the argument to follow, which is modeled on the two arguments of which I have just reminded us, is a slower and easier-to-follow version of Millgram, 2009,

4

With these model arguments before us, let's turn back to David Lewis's modal reductionism. Take that last item in our short list of sample modal claims: If you *were* King Mithridates, you *would have* foiled your enemies' evil plans. On Lewis's way of rendering the content of that counterfactual, it comes to something along these lines: In the nearest possible worlds in which Mithridates is your counterpart, he foils their plans. So to assess the truth of the sentence, we notionally travel out to the nearest worlds in which you are King Mithridates (in which he's your counterpart), and if, in those worlds, he foils his enemies, it comes out true.

Now scratching the surface of even that quite minimal description reveals a great many embedded modal facts. For instance, a king is someone such that, were he to issue a command to his ministers, it would be (*ceteris paribus*) obeyed; this is, after all, constitutive of being a king, or anyway was so until the advent of merely ornamental royalty. A king is someone such that, were he to die, one of his children would succeed him. A king is someone such that, were he to wear a ridiculous garment, no one would dare to comment on it. —*Are* these really necessary? Well, if too many counterfactuals like these are not true of you, we'll start to wonder whether you really are a king.[20]

On the modal reductionist account, paraphrasing away these further counterfactuals means picking out the counterparts of those counterfactual Mithridates in still further possible worlds—possible worlds that are near to the ones they occupy. To check the truth of the *first* counterfactual, we notionally travel out to the nearest possible world(s) in which you are Mithridates, and see if you foil your enemies; but to check whether he really *is* the counterpart required by the antecedent of the conditional (among other things, whether he really is a king), we have to consider the truth of *another* counterfactual (his orders would be obeyed), and so we notionally travel out from some Mithridates-containing world to the nearest possible world(s) to it in which your counterpart's counterpart gives an order, and

sec. 11.5.

[20] We're using Lewis to think through our more general concerns, but alert readers may be worrying that perhaps the argument we are embarking on has too narrowly defined a target. On the views of other theorists about modality, we do not need to identify our counterparts in other possible worlds; rather, we simply stipulate that it is *you* we are considering. (See Kripke, 1980, and Kaplan, 1979, for background.) So notice that we can concede the point without affecting the way the argument plays out: the example would work equally well if the antecedent of the counterfactual was, "If you were a king. . . "; the problem, either way, is determining whether you have what it takes to be a king.

determine whether it is obeyed; also, we travel out to the nearest world(s) in which your counterpart's counterpart dies, and we determine who succeeds him... and so on. (See Figure 1 for help visualizing the spheres of 'nearby' possible worlds, and the ways they are positioned in the natural representation of the example; evidently, this would be an all-too-appropriate occasion to revive the word 'epicycles'.)

But who are *those* counterparts? I adapted the counterfactual whose content we are trying to paraphrase from a bit of Housman's *A Shropshire Lad*, so allow me that Mithridates' counterparts have to be, at any rate, human beings: surely that's how Housman would have meant it! But now, scratching the surface of a human being yields just as many modal facts as scratching the surface of a king. A human being is something that wouldn't vanish if you were to breathe on it. (Things that look like people but vanish when you breathe on them are *ghosts*.) If the putative human walks and talks normally, then it has to be true, if it *is* a human being, that, if you asked him any of a great number of unexpected questions, a suitable answer would come to his mind most of the time. (In other words, he has dispositional mental states; these will become important in due course.) So we owe a possible-worlds paraphrase for these modal facts as well.

When you try to paraphrase ordinary modal talk into possible-worlds talk, using Lewis's recipe for doing that, you find yourself marching down a regress, and one that looks very similar to the regress that refuted phenomenalism. But let's make the problem that Lewis is facing conform fully to the shape of our historical paradigms. Bracketing worries that maybe, even in principle, there's no such thing as a modally thin concept or description or object, suppose we *have* such a description—not 'king', but as close as we can get without building in unwanted counterfactuals—which we use to pick out objects in nearby possible worlds occupying a suitably kingly role. Then the counterfactual we started out with will have to be rendered roughly along the following lines: objects in nearby possible worlds that serve as your counterparts and also satisfy this description (they spend time in a throne room, carry a sceptre, and so on), foil their enemies... *unless* (and here comes the extended *ceteris paribus* clause) they wouldn't be obeyed if they gave orders, or they would be laughed at if they were to wear a ridiculous garment, or anyway, *too many* conditions like the foregoing hold; or they would vanish if breathed on, or they have no dispositional mental states, or... (The standard logical behavior of *ceteris paribus* clauses is on display here: it is obvious that the exceptions are not going to run out, no matter how far you extend the list.)

The force of the *ceteris paribus* clause is *not* that if a king were, e.g.,

disposed to vanish if breathed upon, he wouldn't manage to foil his enemies. Rather, it excludes these cases from the scope of the claim, so as to make it match the sense of the ordinary assertion: someone who says that a king would foil his enemies does not mean to commit himself one way or the other as to what happens in such oddball cases. If we are to track, even roughly, the semantic intentions of ordinary speakers, the paraphrases that are the cash value of the modal reduction must contain such *ceteris paribus* clauses. These deploy the very vocabulary—modal vocabulary, this time—that the reductionist is committed to eliminating. The uneliminated modal freight is not just modally thin possibility and necessity, but counterfactual conditionals (which is, to remind you, why the argument to this point is not a recapitulation of previous objections to Lewis's reductionism). Once again, the reductionist promise, this time of possible-world treatments of modality, turns out to be empty. Lewis's modal reductionism runs aground on the very same argument that defeated some of its prominent twentieth-century reductionist predecessors.

5

The argument is straightforward enough, but the lessons about modality that I am after will emerge only from the back-and-forth of objections and replies, and let's start in on those now.

Recall that the proposed reduction of counterfactuals invokes a similarity space in which we are to imagine the possible worlds as embedded: the more similar they are, the closer. The enterprise is to reduce modal statements to configurations of possible worlds, and the worlds' similarity relations surely count as an aspect of their configurations. So if the similarity relations between possible worlds can be made to do the work of *ceteris paribus* clauses, we can drop the redundant clauses, in which case the reduction goes through as is, and the argument I have been developing against it fails.

Let's spell out the objection a little more slowly, and while we're at it, let's vary the example, since I observed earlier on that there is always *more* packed into a *ceteris paribus* clause: it is not a human if, were it to hear the code word, it would transform itself back into a bug-eyed monster and return to the mother ship. Mithridates' nearest counterparts, we pointed out, had better be human beings, but 'human' is a modally thick concept. If the *ceteris paribus* clauses that exclude such possibilities contain modally thick vocabulary (as does this one), the reduction has not succeeded.

However, the objection runs, Mithridates-like chunks—I'm putting it

this way so as not to beg any questions about counterparthood—of possible worlds similar to this one don't transform themselves into bug-eyed monsters; that in itself would make such a world very strange, from our point of view. And so worlds close to (which again, means: similar to) those worlds also do not contain Mithridates-like chunks that transform themselves into bug-eyed monsters. After all, since Mithridates' counterparts don't transform themselves into BEMs, would-be counterparts of theirs who did would be quite dissimilar to *them* (leaving aside for the moment worries turning on the near-transitivity of 'similar'). So, since the nearby worlds *are* the ones in which Mithridates' apparent counterparts *do* have the right modal properties, we can drop the clauses of the reduction whose point is to add that guarantee. And so the reduction does go through after all: thus the objection we are canvassing on Lewis's behalf.

To address it, we need to stop and think for a moment about Lewis's conception of a similarity ordering over possible worlds. There are two relevant constructions we might put on it: On the first, similarity would be an objective or metaphysical relation, having to do with patterns of what Lewis called *universals*: the natural properties picked out by our deepest theory of the world (which will be, if materialism is true, "something not too different from present-day physics, though presumably somewhat improved"). Alternatively, possible worlds could count as more or less similar in virtue of whether the person uttering the counterfactual in question takes them to be; that is, the similarity relations over possible worlds could be a matter of the similarity judgments attributable to one or another ordinary speaker.[21]

Obviously, what you are committed to reducing depends on what philosophical work the reduction is supposed to do for you. Here we are considering a reduction motivated by the worry that you do not understand your *would'ves* and *could'ves* and *might've*s (recall that I am considering the specifically metaphysical motivations for such a reduction elsewhere), and on the assumption that you are an ordinary speaker of an ordinary natural language, that means the reduction has to capture the content of

[21]For discussion of universals, see Lewis, 1999, ch. 1; the quote is at p. 37. Divers, 2002, p. 123, is an example of a philosopher taking it for granted that appeal to ordinary judgments is the only alternative appropriate for the job it is to do.

Now, there may be convergence between the two prongs of the fork; for instance, Lewis appropriated Davidson's uses of the Principle of Charity, and in particular held that the naturalness of an interpretation (of someone's psychology or utterances) constrained its eligibility. (I.e., we interpret someone as meaning *green* by "green," rather than *grue*, because green is a more natural property than grue; see Lewis, 1983–1986, vol. 1, ch. 8, Lewis, 1999, pp. 45–55.) So we should not assume that the dilemma is a clean choice, and I am constructing the argument to follow with that in mind.

ordinary speakers' ordinary counterfactuals about ordinary objects.[22] Now of course people can be wrong (indeed, very, very wrong) about what would have happened, if... But treating the physics-style universals as providing the underlying measure of comparative similarity makes ordinary speakers out to be wrong in the wrong ways; the systematic errors attributed to them by putting an objective or metaphysical construction on similarity shows that their semantic intentions are not being tracked. For instance, we allowed that ordinary speakers do not mean to commit themselves to what would happen to a Mithridates who was, in virtue of the modal facts, not really a human being but a bug-eyed monster in disguise. A world in which something that looks like Mithridates turns into a bug-eyed monster and returns to the mother ship (or, for that matter, is not obeyed by ministers, or is laughed at for wearing ridiculous garments...) might well be quite similar to ours at the level of basic physics.[23] Since ordinary speakers mean to exclude these cases from the range of commitments they assume when they advance our sample counterfactual (on the basis of their similarity judgments: space aliens are just too *weird*), universals are not suitable for capturing the content of their ordinary counterfactuals.

To make the similarities that drive the ordinary counterfactuals out to be simply *physical* similarities—as a way of showing the similarity metric to be a matter of what is objectively, willy-nilly *there*, as opposed to something derived from what's resident in speakers' psychologies—requires that a big difference, as far as some counterfactual is concerned, be a big *physical* difference. But that's not the way it normally goes. As some of the Anscombians have recently emphasized, in physics there's no answer to the question, What comes next? when it is asked in the register of natural history. (What comes next depends on whether the flower we are examining is, say, trampled by a passerby; it is mere first-this-then-that.) Whereas in biological thought, and for that matter when we are considering intentional action, the question, asked in that distinctive register, is in place. (Next, the buds unfold into small pink blossoms.) In a similar vein, a character in a recently-mainstreamed graphic novel points out that "a live body and a

[22] A popular nonrealist view was that possible worlds were one form or another of representation; one of Lewis's arguments for his own view was that such representations couldn't capture the truisms of ordinary modal discourse, as in his treatment of 'ersatzism' (1986, ch. 3). But if that inability rules out ersatzism, then, by parity of argument, it's not an option, for Lewis anyway, to opt for the possible-worlds dialect and what can be represented in it, and just give up on the aspects of ordinary modal discourse that it can't accommodate: tracking ordinary users' semantic intentions matters.

[23] Allowing for the sake of the argument that 'basic physics' is counterfactual-free, and yes, those are scare quotes. Why? Take a look at Wilson, 2006.

dead body contain the same number of particles." A big biological difference can be a very little physical difference, and, conversely, a big physical difference can amount to, say, a very small economic difference.[24]

We are not quite done with the topic, and I will return to it below, when the time comes to consider supervenience as an alternative to reductions. In the meantime, let's agree to require that when we unpack the black box of comparative similarity, its inner workings must made out in terms of the psychologies of ordinary speakers: what do *they* take to be more or less similar, when they are considering alternative ways things might have been?

Human psychologies are small-finite, and this has as a consequence that judgments of the comparative similarity of whole possible worlds must—with unimportant exceptions—be constructed from reactions to local features of those worlds. The reason is not just that most possible worlds are too complicated for any human mind to survey adequately. Because there's not enough cognition to go around, we generally have to be choosy about where we invest thought and deliberation. Human beings' practical perspectives on the world are quite local; consequently, our well-considered judgments are also, almost without exception, quite local.

Let's distinguish *thoughtful* from *thoughtless* judgments of comparative similarity. We are interested in counterfactuals because we rely on them so heavily in our intellectual and practical lives, and we are not unreasonable to do so. If judgments of similarity underwrite our counterfactuals, then a

[24]Thompson, 2008, p. 41, Moore *et al.*, 2005.

Here's another way to say it: of course a world in which one encounters bug-eyed monsters in disguise must be, intuitively, somewhat different from our own, and maybe even very different in ways that one or another science would recognize. But the task we are considering on Lewis's behalf is that of coming up with a counterfactual-free way of saying how to count those differences, in order to show that counting differences that way preempts the problem posed by *ceteris paribus* clauses for his proposed reduction. If the similarities and differences are made out at the level of special sciences like cultural anthropology (a science that might have a good deal to say about what makes aliens *alien*), it is very hard to believe that they prove free of counterfactuals. If they are nonetheless to be made out at the level of physics, then we need a way of rendering similarities and differences that appear in the descriptions of, e.g., cultural anthropology into the vocabulary of physics. Recall that I began the paper by reminding us that one clear lesson of twentieth-century metaphysics is that such reductions don't work.

Now a typical response on the part of Lewisians is to appeal to the role of context—especially, conversational context—in foregrounding *some* physical dimensions of similarity. However, then the work of selecting the features in virtue of which we judge worlds to be more or less similar is taken up by ingredients which are either psychological, or logically on a par with psychological facts. This is to transfer us to the second horn of the dilemma we are examining. (There are further difficulties with appeals to conversational context; I will register one of them in note 36; for another, see Millgram, 2009, sec. 7.7.)

reliable counterfactual must be tied to well-considered similarity judgments. Counterfactuals whose contents are given by judgments of similarity that have had no thought put into them are worthless. (As the computer scientists say, *GIGO*: Garbage In, Garbage Out.) So we can disregard merely thoughtless judgments of similarity.[25]

Possible worlds are overall or global ways things might be (or might have been). Counterfactuals are enormously important for real life: we make decisions on the basis of our opinions about what would have happened, if... And for that reason, we (often) do our very best to be thoughtful about our counterfactuals, and to get them right. So whatever opinions underwrite those counterfactuals must be commensurately deliberate, that is, we must be about as thoughtful about them as about the counterfactuals themselves. With very rare exceptions, we do not bother having thoughtful opinions about the comparative similarity of global or overall possibilites.

We can confirm this assessment by considering one of those exceptions that prove the rule: physicalism, for present purposes the view that everything is *really just* configurations of physical objects and properties. Take the version of it that amounts to a supervenience claim: if the nonphysical objects and properties were different, the physical objects and properties would be different, too. One way to recast this sort of physicalism in Lewis's conceptual apparatus is this: possible worlds in which nonphysical similarities and differences are not tightly tied to physical similarities and differences are extremely different from our own world. In other words, physicalism itself amounts to a comparative similarity judgment that takes whole possible worlds as its objects, and one that has had thought put into it by physicalist philosophers. But now that we have an example of an actual similarity judgment that takes entire possible worlds as its object, it is obvious how *unusual* this sort of assessment is. I don't mean that physicalism is a minority view among nonphilosophers, but rather that most people (philosophers included) hardly ever invest any thought at all in similarity judgments of this generic logical type.[26]

[25] On occasion, Lewis appealed to our willingness to produce thoughtless opinions about the similarity of items—such as cities—that are too complicated to understand; if cities, he was suggesting, why not possible worlds (1973, p. 92)? This was a mistake on his part, because it entails a violation of the conservation of intellectual effort; you can't get usable, reliable counterfactuals out of thoughtless, off the cuff answers to questions like, "Which city is more similar to Seattle, San Francisco or Portland?"

[26] This is a good point to field an objection that has no doubt occurred to the alert reader: that the counterfactuals true of Mithridates are true of him in virtue of his modally flat, this-worldly, *physical* properties. If it *were* true of someone that he might morph into a bug-eyed monster on receipt of the signal, surely features of his anatomy that account for

However, we do work up thoughtful opinions about more local aspects or features of alternative possibilities. So if Lewis's semantics for counterfactuals deploys the psychological resources we have available, the global comparisons must be assembled out of the local ones.

There is a second observation about those local judgments of comparative similarity to be taken into account. Similarity has the role, in Lewis's account, of what I earlier called a black box; for the reduction to succeed, we would need to show that the black box does not conceal modal notions on which we are tacitly relying. Now, because our thoughtful local judgments of counterfactual similarity are cognitively expensive—we are, again, only interested in similarity judgments that have had thought put into them—and because there is only so much attention and deliberation to go around, we generally form such judgments only when their objects are of interest to us. But we are creatures who live modally saturated lives, and almost all of the objects of our attention are counterfactually thick. Almost everything we care about (or worry about, or strive for) has a usually elaborate counterfactual aspect. We've already seen a couple of examples: many people admire royalty, and whether someone is royalty is largely a matter of what counterfactuals are true of him. Being a person is largely a matter of what counterfactuals are true of you. And it is easy to continue in this vein: People move to big cities because of the cultural opportunities—the things they *could* do, which they know they won't have time to do. People are distressed or joyful because of what *almost* happened to them. People care as much or more about whether they *can* be victims of violent crime as whether they are *actually* the victims of violent crime.[27] So almost all of the local similarity judgments we are considering will have as their objects not modally or counterfactually flat properties or states of affairs, but rather, objects or properties or options characterized in just the modally-laden way that the reduction under consideration is committed to eliminating.[28]

it will turn up in the autopsy video.

Suppose you don't want to make physicalism, as just introduced, necessary: after all, most of us think that there might be someone, or something, who walked and talked just like we do, and morphed into other shapes for good measure, but who was hollow inside; it's merely that such a possibility is very strange. Then the coordination of at-a-world physical and modal properties is local: that's what it's like *around here* (in the big and variegated possible world universe, as ordered by similarity). But the reduction we are being offered is not meant to work only locally.

[27]Ruth Chang alerted me to the emotional importance of near-misses; the point about how important inviolability is to people comes from Nagel, 2007, which addresses itself to the centrality of the modal good in the design of political institutions. I am told that Jerry Fodor has also noticed the point about the cultural opportunities.

[28]The worry that a Lewisian similiarity ordering will turn out to depend on the very

In the way of thinking we are working our way through, modal and counterfactual claims are to be understood by way of a picture in which the ways things might have been are *global* (they are ways *everything* might have been, all at once) and *modally* or *counterfactually flat*. The similarity ordering that is meant to defuse the argument we constructed in sec. 4 must thus have these global, flat ways things might have been—possible worlds—as its objects. The judgments of similarity we actually have on hand are, we saw, local rather than global, and have modally and counterfactually thick intentional objects. So if Lewis's reduction is to work, there must be a way of assembling judgments of the latter sort into similarity orderings of the former sort. In a moment, I will proceed to consider whether this is possible in principle, but first I want to register a bit of nuance.

I am trying throughout to stay as far as I can within the spirit of Lewis's proposal, once again, because I think that in doing so I can hit a target of more general interest; the reason I think so is that similarity mostly functions as a placeholder, a representative for whatever does the job of accounting for the truth values of counterfactuals. When he entertained the notion of similarity orderings, Lewis was ambivalent about the idea of writing down such things. He took the similarities in question to be vague, incomplete, and context-sensitive, and even suggested that any way of making them completely precise might thereby misrepresent them.[29] Both our own objectives and Lewis's hedges mean there's a delicate balancing act to take note of here. Lewis took 'similarity' to be a theoretical notion, one that we can adjust to match our control of counterfactual conditionals. This means that you can't always take responses based on eyeballing to pick out its contours. On the other hand, the relation is used to control our assessment of counterfactuals, and so it must be deployed. That means that we must have a pretty good idea of what is, in the relevant sense, more similar to so-and-so than what else, in most situations of ordinary concern. I am going to rely on this point about our competence in what follows, as I introduce a handful of very straightforward claims.

---

sorts of modal judgments that it is trying to explain appears as early as Fine, 1975, at p. 455.

[29] See Lewis, 1986, p. 21, and Lewis, 1983–1986, vol. 2, pp. 181f (on "tailoring"); Lewis, 1986, p. 254. "Imprecise [comparative similarity] may be; but that is all to the good. Counterfactuals are imprecise, too. Two imprecise concepts may be rigidly fastened to one another, swaying together rather than separately, and we can hope to be precise about their connection" (1983–1986, vol. 2, p. 6). See also a remark on "the questionable assumption that similarity of worlds admits somehow of numerical measurement" (p. 12), and related discussion at p. 163, as well as at Lewis, 1973, pp. 50–52, 67.

6

In sec. 3, I described Putnam's argument against behaviorism as another instance of the argument pattern that we deployed against the possible-worlds reduction of modality. Lewis paid close attention to that argument, and his own views in philosophy of mind were constructed as a response; the technique he developed has been adapted by his followers into a popular and systematic approach to the problems of metaphysics (the so-called *Canberra Plan*). Although he seems never to have considered it himself, the Canberra Plan is the best way Lewis had to take up the task we have just outlined, and (once again because I think can learn a number of more general lessons about modality from it) I will first describe the Canberra Plan, and how it might be adapted to the construction of counterfactual-free similarity orderings over possible worlds. I will consider two related obstacles to producing such constructions. Then I will take time out to draw morals about the logic and function of *ceteris paribus* clauses, both in counterfactual contexts and in general.

Recall that Putnam's objection to logical behaviorism exploited the idea that connections between mental states and behavior are mediated by other mental states. (What you do, when you want something, depends on what else you want.) Lewis was happy to allow that, and his variant of functionalism—'analytical functionalism', to distinguish it from Putnam's computation-oriented formulation of functionalism—explicitly accommodated that idea.[30] The device he used required, first, collecting the platitudes of 'folk psychology', then conjoining them into a single, very long sentence, and finally replacing the psychological vocabulary with variables bound by existential quantifiers. Doing that ('ramsifying') gives you a theory in which platitudes like

6. If someone is in pain, he tends to scream

   and

7. If someone has a very strong desire not to scream, he tends not to

reappear as segments of that very long sentence, and look something like this:

8. $(\exists x)(\exists y)$ ... if someone is in x, then he tends to scream, & if someone has y, then he tends not to scream, & if someone is in x and has y then...

[30] Lewis, 1999, ch. 16, Lewis, 1983–1986, vol. 1, chs. 6, 7, 9.

The idea is to take all the theoretical relations at once, and treat the theoretical entities as the occupiers of slots in the matrix that the relations jointly constitute. A theoretical entity is picked out, more or less, as the occupier of slot number n in the theory, and rearranging these characterizations into definite descriptions allows you, in principle, to eliminate the problematic theoretical vocabulary, which means that the technique lives up to the formal requirements of a reduction. In fact, the reductive paraphrase is never executed, but these multiple-slot definite descriptions allow you to identify the innocuous (in this way of thinking, the purely physical) states or properties which as a matter of fact occupy the slots in the theory: folk psychology may be full of beliefs and desires, peculiar mental states on which (you are concerned) you may not have a satisfactory philosophical grip; but once you know that, in human beings, only such and such neurological states are related to each other and to perceptual inputs and behavioral outputs in the way the ramsified folk theory says, you can pick these out as what the beliefs and desires in fact are.

The 'woulds' and 'coulds' with which we started cannot be smoothly inserted into the template of analytical functionalism, but assembling a theory of similarity in the way the model suggests is more promising—though in order to explore the application of the Canberra Plan to Lewis's modal reduction, we are going to have to allow ourselves to depart from the letter of what is now a widely applied recipe. The approach is evidently a way—evidently, the *only* way—to make a similarity ordering over possible worlds serve Lewis's reduction of modality, and here's what it would take: We collect the psychologically available judgments of similarity from the 'folk'; we conjoin these to obtain the folk theory of (counterfactually-relevant) similarity. We elicit from this theory—perhaps via the recipe's step of replacing theoretical terms with bound variables, or perhaps in some other way—a matrix of similarity relations between the different objects, properties, states of affairs and so on. (Recall that these objects, states of affairs, etc., for the most part bear modally and counterfactually thick characterizations.) Finally, we identify the modally or counterfactually flat *physical* properties, states of affairs, and so on that in fact occupy those slots in the matrix. We are then to reconstruct a reduction-compatible rendering of the similarity ordering over possible worlds using only these modally or counterfactually flat elements.

It is hard to believe that this is a program we will actually get around to implementing. But is it possible in principle? If it is, then perhaps, even if we cannot exhibit the Lewisian reduction of modality to possible worlds, such a reduction can be known to be possible in principle. And if it can,

perhaps almost as much philosophical work can be done with that conclusion as if we had the reduction on hand. As the reader no doubt expects, I am about to argue that it is not possible, even in principle.

7

A von Neumann-Morgenstern utility function is way of summarizing a particular agent's preferences, and it is possible to do so provided the agent's preferences satisfy a handful of actually quite demanding consistency requirements.[31] Over the past decades, an objection has crystallized to many of the uses made of expected utility theory: that it is almost inevitable that human beings have both *too few* and *too many* preferences to have utility functions. That is, as a matter of psychological fact, a human being's preferences are far too sparse to induce a utility function; moreover, they are not sufficiently consistent with one another to induce a utility function (or a usable approximation to one, or even a usable range of them). I am about to use this train of thought as a model for considering what can be had by way of similarity orderings over possible worlds. To anticipate, if our comparative local similarity judgments are too sparse to be assembled into comparative similarity judgments whose objects are global or overall alternative possibilities, and if they are mutually inconsistent in ways which prevent them from being so assembled, that will amount to a decisive criticism of the Canberra-Plan approach to similarity orderings over possible worlds. A reminder: we are going to be as interested in *why* the approach fails as in the brute fact that it does.[32]

First, we have *too few* local similarity judgments. To say what I mean by that, I first need to remind you of the contrast, which we invoked in passing in considering Mithridates' modal properties, between occurrent and dispositional mental states. Occurrent states are what's explicitly before your mind, and dispositional states are those that would come to mind when suitably prompted: you're not always thinking, "My name is _____" (so

[31] See Mandler, 2001, for an overview. Here and below, 'consistency' is used in the technical sense in which it figures into such discussions, and without any endorsement on my part of the implicit suggestion that these are conditions our preferences *ought* to satisfy, and that 'inconsistency', in this sense, is a fault and grounds for complaint. (For an argument to the contrary, see Millgram, 2005, ch. 10.)

[32] Lewis himself used to assume that people have (or approximately have) utility functions; that was perhaps an allowable error when he was writing, but, now that Daniel Kahneman has been given his Nobel Prize, is so no longer. Over and above Kahneman's work with his late collaborator, Amos Tversky (1982), representative contributions to this body of work include Shafir, 1993, and Ainslie, 1992.

it's not an occurrent belief), but if someone were to ask you your name, the answer would be immediately forthcoming (so it *is* a dispositional belief). Most of anyone's mental states are dispositional, and because occurrent psychological states are so few and far between, the judgments to which the specification of a similarity relation over possible worlds must help itself will inevitably be by and large dispositional. Your *actual* similarity judgments (the ones that, in Lewis's way of thinking, you have in *this* possible world) are quite sparse indeed. A moment ago, were you actively thinking of any of them at all?

A dispositional psychological state is a state that you have counterfactually.[33] Consequently, unpacking the black box of similarity in Lewis's reduction means bringing to bear judgments of local similarity that ordinary speakers *would* have in one or another sort of counterfactual circumstance. Let's allow ourselves psychological resources from other possible worlds; after all, we can only assemble a usable comparative similarity ordering if we are willing to do so. Again, those psychological resources are your counterfactual judgments about local aspects of similarity. But now observe the vicious circle we are facing. In Lewis's way of thinking, the judgments you would have are those that your *nearest* (i.e., most similar) counterparts have, in suitably specified circumstances. So in order to specify the similarity relations over possible worlds, we need to determine who your nearest counterparts are. But what counts as near and far in the possible world similarity space is a matter of what similarity ordering is chosen. So to determine who your nearest counterparts are, you must first have the similarity relations—which is what we started off asking about in the first place.[34]

[33]Here I am going to ignore the complexities in unpacking dispositions into counterfactuals that travel under the heading of the 'conditional fallacy'; the problem is introduced in Shope, 1978.

[34]Here the argument turns on which of your counterparts is 'nearest', and not on who your counterparts *are*. But in Lewis's own scheme of things, that's up for grabs, too: who your counterparts are, in a particular possible world, is a matter of who is intrinsically similar to you, and occupies a suitably similar location in that world. In Lewis's picture, it's a mistake to think of your modally-extended self as stably composed of your modally thin counterparts, in the way that your temporally extended self is stably composed of momentary time slices. Lewis thought of the, as it were, modality slices of a person in different possible worlds being slices of the *same* person as on a par with the fact that certain stretches of asphalt are all parts of I-5. There's no deep metaphysical fact underlying the 'unity over space' of I-5, and if the highway commission were to decide to rename part of I-5 to be the David Lewis Memorial Freeway, they wouldn't be making a metaphysical mistake. I am not pressing this problem because Kripkeans will be willing to treat your reidentification in other possible worlds as a primitive, and we are after points

Here's an illustration of the problem. A former girlfriend used to dye her hair a very dramatic blonde. Now, she could have dyed it green, and she could also have dyed it purple. I never considered which state of affairs would be more similar to the actual state of affairs, but I'm pretty sure that, had I been prompted, I would have been able to express an opinion. But now, which opinion is the one that ought to get factored into a reconstruction of my theory of counterfactual similarity? If the counterpart who thinks that green is more similar is closer to the actual world, then *this* opinion belongs among the raw materials of the Canberra-Plan theory. But if the counterpart who thinks that blue is more similar is closer to the actual world, then *that* opinion belongs among the raw materials. It is quite plausible (though without seeing the details, it's hard to be sure in any particular case) that the differing inputs will make a difference to which counterpart *is* considered closer: after all, if dying her hair blue does count as more similar to the way things are than dying her hair green, then a counterpart who thinks otherwise is *surprisingly* mistaken (and so, *ceteris paribus*, farther away)—and, of course, vice versa.

Occurent judgments of the relative similarities of alternative possibilities are quite sparse. We need to leave to one side counterfactual judgments whose deployment involves a vicious circularity: meaning, all those would-be judgments of local similarity that are not tightly enough anchored to the occurrent judgments to prevent the sort of problem we just saw from arising. And so when we try to unpack the appeal to similarity, we find that we do not have enough in the way of psychological materials to exhibit the inner workings of the black box, and to show that the reduction works: there are too few judgments of counterfactual similarity available, and the black box, it turns out, is almost entirely empty. We are in principle not in a position to show that similarity can do the work of those *ceteris paribus* clauses.

8

Now let's imagine that the materials that have just proved to be out of reach are nonetheless at hand. Again, those materials are similarity judgments about local features of possible worlds, rather than assessments of possible worlds in their entirety. Remember those von Neumann-Morgenstern utility functions, and remember that it is now a commonplace that individual preferences are not regimented into those patterns: human beings rarely if ever *have* utility functions. An easy explanation, though likely not the only one,

in our treatment of Lewis that will travel.

is that imposing decision-theoretic consistency on independently-generated preferences is too hard a cognitive task. When you stop and form a preference over two objects of choice, you have to have a special reason to check if it is decision-theoretically consistent with other pairwise preferences you have formed on other occasions. It would be impossible, or next to impossible, to check that the new preference is consistent with all the preferences you have already adopted, or even with most of them. So, again, unless there is a special reason to do so, there will be no reason to expect consistency from pairwise preferences formed on different occasions.

That explanation can serve as a model for the argument at hand. If similarity judgments are generated one by one, to address local concerns, and there is no systematic and concerted effort to render them consistent, it would be an unbelievable coincidence if they *were* consistent. This is certainly true of generic (rather than counterfactual-specific) judgments of similarity. Consider that, during the Cold War, Hungary was like the Soviet Union, but the Soviet Union was not like Hungary; or again, a coffee shop in Nashville formerly displayed a cinnamon bun that looked, its advocates claimed miraculously, like Mother Teresa, but Mother Teresa did not look like the "Nun Bun".[35] I expect that human beings are rarely if ever in a position to perform a sufficiently ambitious consistency check on their independently formed local similarity judgments; it is simply beyond their (our) cognitive capabilities.[36]

[35]I'm grateful to Dedre Gentner for the first of these examples. Medin *et al.*, 1993, collects evidence that independent processes produce uncoordinated similarity judgments; the Hungary example is a variation on an instance reported by Amos Tversky (p. 259). Gentner and Rattermann, 1991, documents some of the ways in which judgments of similarity are tied to developmental stages.

[36]For a complaint about Lewisian similarity orderings of possible worlds that is plausibly a side-effect of the cognitive limitations at which I am gesturing, see Preuss, 2007.

There is now a standard way of responding to examples like those I have just given: to insist that similarity is context-sensitive, and that the context changed, mid-sentence, in both examples. And some philosophers have developed the habit of gesturing at the moment-to-moment variability of the similarity space, when faced with one or another problem, as though the problems were thereby solved. But taking the relevant similarity relations over possible worlds to be fluid and context-sensitive makes it *harder* to show that they do their job, not easier. Recall the problem of the previous section, that there aren't enough in the way of available materials to reconstruct enough of a similarity ordering over possible worlds to save Lewis's reduction. If there are many different sets of similarity relations in play, and we switch off between them, moment to moment, then the materials available for reconstructing any one overall similarity ordering are vastly fewer. In other words, the appeal to context, invoked as a way of addressing inconsistencies, makes the sparseness problem *worse*: if before we did not have enough for a single, stable similarity ordering, we will hardly have enough for the many ephemeral context-dependent orderings.

If that is correct, it is not just that we do not have enough in the way of raw materials to reconstruct the global similarity ordering that Lewis's semantics for counterfactuals requires; we also have *too much*. It is evidently overdetermined that Lewis is not in a position to show that a similarity metric can do the work of the *ceteris paribus* clauses that frustrate his reduction of counterfactuals to possible worlds. If he is not, the objection we were considering lapses, and, as anticipated, the reduction fails. But there is a more interesting lesson to take home from our guided tour of Lewisian arcana, having to do with the deeper reasons that the local judgments aren't suitable raw materials for the global ones, and I now turn to that.

9

The argument we have just completed emphasized that our judgments of local, counterfactually-relevant similarity are typically mutually independent; it tells us that in our modal cognition we get by with problem- or topic-specific sketches of the modally important features of the circumstances, and that we do not ordinarily assemble these sketches into a global, consistent and counterfactual-free Big Picture of the modal facts around us. (We cannot come to have that sort of a theory; the Canberra Plan reads an analysis of a given subject matter off just that sort of theory; that was why the Plan turned out to be unusable.) But how do we manage to navigate using these partial and routinely jointly inconsistent sketches? That is too large a question to take on here, but we can consider a runup to it. If this is how we use our local similarity judgments, we should expect to find them prepared, so to speak, for the use they get. Do we?

Return to the example of two sections back: would dying her hair purple have been more or less similar to the way things actually were than dying her hair green? Let's suppose that I judge the latter to be more similar: if I do, that opinion is advanced as correct only *ceteris paribus*. After all, if she were dying her hair green because her FSB spymasters had instructed her to poison a wealthy Russian emigre with exotic radioactive materials—well, that would make it much *less* similar. If I am seeing the territory correctly, such implicit *ceteris paribus* clauses are there to allow for the friction between independently generated judgments of counterfactual similarity, and constitute, as it were, logical preparation for working with otherwise inconsistent materials. The price of thus softening (or allowing us to paper over)

It is also worth reminding ourselves that not all incoherences can be conjured away by appealing to shifts in context (a point emphasized by Lewis, 1973, p. 13).

these inconsistencies is that opinions that embed *ceteris paribus* clauses (at any rate, the type of *ceteris paribus* clause capable of serving this logical function) are not well-behaved under conjunction: that's precisely what it takes to make the inconsistencies go away. And this gives us another way of saying why the Canberra Plan won't work here: the first couple of steps in ramsifying a theory are to collect all of the claims we make about some subject area, and then to conjoin them into a single, long sentence. But you should only be willing to assert the theory (the single, long sentence) if you take conjunction to be truth-preserving, and because our local similarity judgments contain implicit *ceteris paribus* clauses, it isn't.

We were examining local similarity judgments because they were meant to underwrite the behavior of our counterfactuals; let's confirm that the phenomenon surfaces there as well. Consider the following counterfactual: if I had a second car, it would be a Hummer. If that's true, it's true only *ceteris paribus*; if I were an avid UFOlogist, my second car would be a black Cadillac, to mislead the government agents whom UFOlogists believe are persecuting them, and who drive black Cadillacs, into thinking I was one of them.[37] In Lewis's way of thinking, possible worlds in which I am an avid UFOlogist are farther away, i.e., less similar to the actual world, than the nearest world in which I own a second vehicle. That is to say that the initial underlying judgment of similarity (a way things might be in which my second car is a Hummer—this is now an incomplete possibility rather than an entire possible world—is more similar to the actual way things are than a way things might be in which my second car is a Cadillac) is true only *ceteris paribus*: its conjunction with 'I am an avid UFOlogist' comes out untrue.[38]

[37] Mitchell, 1999, p. 229.

[38] There is an older literature focused on what can now see to be a misconception arising out of the assumption that local sketches of the modal territory can be assembled into a global, internally consistent map. Suppose we are evaluating a counterfactual such as, If I had looked in the mirror, I would have seen my own reflection. On the assumption that we live in a deterministic world, the antecedent of the counterfactual requires one of the following three alternatives: a deeply and pervasively different past, different natural laws, or a miracle—a 'jump' whereby I inexplicably come to look in the mirror. And so the problem of how (putting it in Lewisian vocuabulary) to assess the relative distances of such worlds from our own came to seem pressing, and indeed received a great deal of attention.

The *ceteris paribus* clauses attached to local similarity judgments are—with insignificant exceptions—bound to be triggered by any of these alternatives. That is to say, having rendered such a judgment, and having been informed that one of these alternatives is now part of the story, you will retract the judgment of similarity; when it comes to matters such as miracles or systematically different pasts, we become agnostics about similarity

Generally, then, the local similarity judgments we are contemplating must be understood as containing *ceteris paribus* clauses, and this allows us to account for a claim about the logic of *ceteris paribus* clauses that I made in passing earlier on. The function of these *ceteris paribus* clauses is to anticipate and accommodate the potential conflicts among independently formed judgments; because one does not normally stop to survey one's other views about counterfactually-relevant similarity before forming a particular judgment about it, that judgment might later on have to coexist in your intellectual world with just about *anything*. What comes under the heading of *anything*? Since Nelson Goodman, it has been a truism that there are infinitely many ways that any two things can be similar, and infinitely many ways in which any two things can be dissimilar. So the *ceteris paribus* clauses implicit in counterfactual-supporting local similarity judgments will exhibit a distinctive sort of open-endedness; there can always be further similarities or dissimilarities which properly suspend the judgment. That just means that there is always, as I remarked, *more* built into a *ceteris paribus* clause.

Knowing what underwrites the logical behavior of such *ceteris paribus* clauses allows us to dispose of two objections to our original argument, which I have found often occur to Lewis's followers. First, on what we might call the statistical conception of *ceteris paribus* clauses, such a clause means: the claim holds with high frequency, or with a small number of exceptions. And this might lead someone to expect that the *ceteris paribus* clauses in the proposed reduction of counterfactual claims to possible-worlds claims can be ignored, or that they will wash out, or that they can be exhausted. (If we go on paraphrasing for long enough, eventually we will run out of exceptions, or at least the exceptions will be *few* enough to be negligible.) Second, someone might defend an application of the Canberra Plan in this way: the idea behind the Plan is to take all the theoretical relations at once; as it was once put to me, Lewis's favorite move was to stuff *all* the relations into the box and close the top over them. Since the *ceteris paribus* clauses are just more relations between the theoretical entities, there's no reason, you might think, why they too can't be integrated into a reduction.

However, you can only collect *all* of a *definite* number of sentences; conjunction is an operation well-defined over finite sets of sentences (and which can, with a little bit of ingenuity, be well-defined over countable sets

and counterfactuals: all bets are off. These puzzles arose in the first place because it was assumed that the local renderings must be glued together into a single 'possible world'; the assumption that we are considering the world as a whole is built into the use made of the premise of determinism. The function of such *ceteris paribus* clauses tells us that we should have known better than to try.

of sentences), but not over *indefinitely* many sentences. We have just seen why it is characteristic of *ceteris paribus* or 'other things equal' clauses that there are always *other* 'other things' to be 'equal'; we can no more count or survey the ways in which there always further things to go wrong than we can count or survey the ways in which objects or states of affairs can resemble or fail to resemble each other. That means that we can't collect all the relations at once; there's no theory, containing explicit renderings of all of those 'other things', to ramsify (i.e., once again, we have explained, from a slightly different angle, why the Canberra Plan is a nonstarter here).

We have now underwritten our earlier willingness to treat *ceteris paribus* clauses as inexhaustible. However, on the Lewisian understanding of *ceteris paribus* clauses, there aren't really indefinitely many things to go wrong; *ceteris paribus* clauses don't essentially contain an ellipsis, or if they do, unpacking those ellipses comes to an end. That understanding is an error about the logic of *ceteris paribus* clauses generally, and we now have a subject-specific explanation for how it is that the *ceteris paribus* clauses which appear in the course of paraphrasing counterfactuals possess the distinctive logical open-endedness on which our argument turned. In fact, as we have just seen, even Lewis's own apparatus, taken together with platitudes about the available dimensions of similarity, commits him to this logical feature of *ceteris paribus* clauses.

10

We still need a reason for enforcing the reductionist demand, one that goes deeper than: Lewis happened to have accepted it.

A claim is not contentful merely because it is phrased in familiar words, and a philosopher's duty is to give his claims content. Reductionists owned up to their duty by promising to translate away the vocabulary they proposed to discard. If a philosopher wants to let go of the reductionist way of giving his claims content, then he must supply a substitute; acting as though his words meant something, when he has not done any of the work required to make them do so, is not an option.[39]

When previous reductionist programs failed, their adherents retreated to claiming that one kind of thing (that they had failed to reduce) *supervened* on the other (the kind of thing they'd failed to reduce it to). Supervenience has been the traditional substitute for reduction, and for a long time, until it

[39]These remarks are meant to address confusions that appear to be quite widespread. A typical instance is Divers, 2002, pp. 28f.

was supplemented by the Canberra Plan, it was the only substitute in general use. When such a move gets made, the picture stays the same: biological objects and facts, say, are really just configurations of physical objects and facts... only without the obligation to say *which* configurations. The move is often accompanied by the announcement that the claim being made is about ontology or metaphysics (more recently, about what the 'truthmakers' are), and is not intended as conceptual analysis. Put in less hifalutin language, although you are giving up on saying what your words had meant, you are still trying to say what they were—collectively—really *about*.

But one way or another, a philosopher offering a claim about what is really what else has to answer the question, "Where's the beef?" The beef which a supervenience theory supplies is the claim that if, say, the biological supervenes on the physical, then the biological facts couldn't be different without the physical facts being different, too.[40] Lewis himself was a fan of supervenience in other domains, and so let's consider whether modal supervenience was available to him as a fallback position. In Lewis's picture, the modal facts are really just facts about the configurations of possible worlds. Even if he couldn't say *which* configurations, he ought to have insisted that the *coulds* and the *woulds* and the *musts* supervene on the facts about the possible worlds: if the modal facts are just a matter of how the possible worlds are configured, you can't change the modal facts without changing the possible-world configuration.

Or was that really an option? Again, the cash value of supervenience is that, if the supervening facts were different, the supervened-on facts would have to be different. But that's a modal claim, and Lewis acknowledged that it was: "we have supervenience when there could be no difference of one sort without differences of another sort. ... Clearly, this 'could' indicates modality. Without the modality we have nothing of interest."[41] Unless the claim can be accounted for using the apparatus of possible worlds, the position is self-refuting. Can it?

What the set of all possible worlds is (and how it's configured in 'similarity space') was not, on Lewis's way of thinking, a contingent matter: it *could* not be different than it is. So Lewis himself was not in a position to so much as articulate the fallback supervenience claim. But we can ask whether, regardless of Lewis's own view of the matter, the position is theoretically viable. Making out counterfactuals about the set of all possible worlds would involve modal realism about other possible super-worlds, each

[40] See Lewis, 1999, pp. 33–39, for complications.
[41] Lewis, 1986, pp. 14f.

of which is a way the set of all possible worlds might have been.[42] Since we are imagining worlds over and above the possible ones, this extension of Lewis's approach amounts to the currently popular idea of supplementing possible worlds with 'impossible worlds'—or rather, with universes of them.[43]

To think intelligently about the counterfactual covariance of modal facts with configurations of possible and impossible worlds, we would need to make sense of counterfactuals such as, "If the configuration of possible worlds had been different, then...," For this, on the approach we are trying to extend, we would need a similarity ordering over the impossible worlds, and over universes of impossible (together perhaps with possible) worlds. But we have just argued that we are not even in a position to work up a usable similarity ordering of *possible* worlds. *A fortiori*, we are not going have the wherewithal to construct the far more demanding similarity ordering. How much does any sane person have in the way of thoughtful opinions about which features of *impossible* worlds make them more or less similar to each other? And how much does anyone have in the way of thoughtful opinions about what features of alternate universes of worlds make them more or less similar to the Lewisian universe of possible worlds? Once again, thoughtful opinions are the product of attention, deliberation and, generally, cognitive work. No one in his right mind has paid any attention at all to such matters; therefore, no one in his right mind has the thoughtful opinions that would underwrite the sort of counterfactuals required to make sense of the supervenience of the modal on possible worlds.

[42] As per Skyrms, 1976, p. 327n.10; the option is one that Lewis explicitly considered and rejected: "it makes no sense to repeat the very method you think has failed, only on a grander scale... There is but one totality of worlds; it is not a world; it could not have been different" (1986, p. 80). "It is futile," he subsequently wrote, "to want the entire system of worlds to satisfy a condition, because it is not contingent what conditions the entire system of worlds does or doesn't satisfy" (1986, p. 125).

Divers, 1999, has attempted to extend and defend Lewis by offering a 'redundancy interpretation' of modal claims about the machinery: statements like "It is possible that there are many worlds" are allowed, but they are flattened down to "There are many worlds." So notice that the way Divers provides of stating supervenience claims about modality disbars them from doing any of the work that we needed supervenience for: we can no longer capture the thought that, were the modal facts to vary, the configuration of the totality of worlds would have varied as well. (Compare, on this point, Divers, 2002, pp. 55–57.) However, Divers, 2002, pp. 208f, explains how some types of reductionism about possible worlds—that is, views on which possible worlds are themselves to be reduced—can accommodate supervenience claims naturally.

[43] Yagisawa, 1988, which gives a very funny two-front argument against Lewis exploring the option of 'impossible worlds'; Lewis did not regard the suggestion as a friendly amendment to his view (1986, p. 7n).

The point is that we can't afford to be casual about the failure of a strict reduction. It is not as though Lewis could have retraced the steps taken by earlier embattled reductionists, and backed off to a weaker but still contentful modal supervenience account. When you back off from insisting on a reduction, you have to replace it with something else, if you're going to end up saying anything at all. When early analytic philosophers did their metaphysics (all the while denying that they were), they spelled out the contents of their claims linguistically, as theses about what could and couldn't be given eliminative paraphrases. More recent philosophers have spelled out the contents of their more modest claims via the Canberra Plan and via supervenience. But these devices are not available, which was why, when Lewis discussed modality, he wrote like an old-style reductionist. If you back off from old-style reductionism, but don't say what you're backing off *to*, then you haven't managed to make contentful claims, and if the Canberra Plan and supervenience are unavailable, it's reductions or nothing.

Let's go back to an option we were considering a while ago, that a Lewisian similarity ordering over possible worlds might be constructed out of the objectively (that is, physically) present properties of those worlds, and without appealing to our own psychologies, but rather, to universals. Can that claim be softened out to: the similarities and dissimilarities we discern among possible worlds *supervene on* (even if they can't be reduced to) their objective, universal-based features? A supervenience claim has it that there's no difference between (say) nonphysically characterized states of affairs unless there's also a physical difference. The point of the move to supervenience is to prescind from telling you *what* the physical difference, in such a case, is. And so it follows that a supervenience theorist can't back up a claim to the effect that a big difference of some nonphysical kind is a *physically* big difference. And so it follows that a supervenience claim cannot do the work needed to save Lewis's account from our initial objection to it.[44]

[44]At this juncture, you might be wondering whether I'm being sufficiently charitable. After all, it is quite normal to assert counterfactual conditionals about matters that could not be otherwise, as when, working a mathematics exercise, I say to myself: suppose that the height of the formula *were* greater than the theorem allows, then this variable *would* take such-and-such a value... (Would Lewis have gone along with this suggestion? We are told that "nothing can depend counterfactually on non-contingent matters. For instance nothing can depend counterfactually on what mathematical objects there are, or on what possibilities there are. Nothing sensible can be said about how our opinions would be different if there were no number seventeen, or if there were no possibility for dragons and unicorns to coexist in a single world. All counterfactuals with impossible antecedents may indeed be vacuously true. But even so, it is seldom sensible to affirm them" (1986,

11

Lewis gestured at a method of converting ordinary modal language (woulds and would haves, coulds, musts, and so on) into a possible-worlds paraphrase, and in this he was typical. He was also attempting to enforce what he thought were our ontological commitments to the possible worlds, and in this he was atypical. There is only so much mileage to be gotten from disabusing us of something that no one else believed anyway, so let's focus on the common ground: pretty much everyone in the business takes it for granted that ordinary modal speech, including counterfactuals, bald claims about possibility and necessity and so on can be converted into the possible-worlds vocabulary.[45] Even those who insist that possible worlds

p. 111).) Let's imagine that Lewis would in the end have to have allowed for some way of construing counterfactuals of this sort. Why can't it (whatever it is) be used to underwrite the counterfactuals about the set of all possible worlds that would allow us to make sense of modal supervenience?

Some way of understanding such counterfactuals there must be. Whatever it is, however, it is not to be made out using the apparatus of possible worlds. So it must be done some Other Way, and that Other Way will ultimately need to be given a philosophical explanation. The Other Way is going to have to be very powerful indeed, if it is going to handle counterfactuals about how the possibilities themselves might have been different. There is no surface difference between those counterfactuals that require treatment via the Other Way, and those that are amenable to Lewis-style possible-worlds renditions; laymen don't distinguish between the counterfactuals of mathematical reasoning, and counterfactuals about their garage work. So why shouldn't we expect that, when we have the account of Other Way counterfactuals, it will handle the phenomena that Lewis's modal reductionism was supposed to handle? In short, our final worry that we are being uncharitable turns out to presuppose a further and distinct account of (anyway certain kinds of) modality, one which we can reasonably expect to make Lewis's own account superfluous.

(Divers, 2002, p. 98, offers a companions-in-guilt response to the complaint that Lewis and his followers cannot handle such counterfactuals: nobody else can explain them either. So notice that this reply is irrelevant to the point I am making here. Notice also that the inability to handle such counterfactuals is an objection to Divers's proprietary treatment of apparently modal claims about the modal machinery (that is, of what he calls 'extraordinary' modal claims, mentioned in note 42, above); since counterfactuals are of a piece with the rest of the machinery, that treatment should extend gracefully to cover 'extraordinary' counterfactuals as well, and it does not.)

[45]Compare Chihara, 1998, pp. 144ff; typically authors who allow that the expressive power of a language with modal primitives and the possible worlds vocabulary can differ think that what you can say in the former, you can say in the latter. E.g., Melia, 1992, argues that the possible-worlds vocabulary allows you to say *more*. And when Lewis, 1986, pp. 10–13, faced up to difficulties in rendering ordinary modal expressions into a regimented modal vocabulary, he resolved the problem thusly:

> If this language of boxes and diamonds proves to be a clumsy instrument

have modality built into them think that the paraphrase is available. Even those who are philosophically unhappy with the possible worlds respond by attempting to paraphrase possible worlds talk away; the assumption implicit in this way of proceeding is that the vocabulary of possible worlds gives us the right expressive power (it captures the content of ordinary modal speech); we just want a different way of getting precisely *that* expressive power. What we have seen, however, is that possible-world renderings don't match the commitments of ordinary statements: as when, in our example, you took on *extra* commitments, to what something which looked a lot like Mithridates, but was really a bug-eyed monster, would do.[46]

No one has a very good philosophical account of modality, and in my view, the habit of talking of about possible worlds is in significant part to blame. When the possible-worlds way of paraphrasing modal vocabulary hit the diaspora of analytic philosophy departments, it was adopted as a distinctive if odd manner of speech—a kind of scattered regional dialect. In the South, they say "you all" (or, in some of the more rural areas, "you'uns"); in the philosophy departments, they came to say things like, "in some possible world, you are a contender." The accent is easy to acquire, and because it was assumed that what you could say in one way, you could say in the other, these philosophers acted as though there really was not much more to it than an accent. But the philosophers (and philosophers-in-training) who did so for the most part had the impression that they thereby *understood* modality—maybe not everything about it, but enough for their purposes. Maybe one didn't know what possible worlds *were*, and maybe there were other issues to argue about, such as whether other possible worlds really *exist*, or whether you could reduce the modality away, but one at any rate

> for talking about matters of essence and potentiality, let it go hang. Use the resources of modal realism *directly* to say what it would mean for Humphrey to be essentially human, or to exist contingently.

In other words, the possible-worlds vocabulary is perhaps more powerful than the language of quantified modal logic, and powerful enough to render ordinary claims about essences, etc. (However, for dissent about the counterpart-inflected vocabulary, see Fara and Williamson, 2005.)

[46] One might wonder whether having to surrender the possible-worlds renderings of counterfactual discourse nonetheless allows us to keep possible-worlds renderings of thin modal vocabulary. Perhaps "Mithridates might have worn a chocolate crown" can still be construed, without change of content, as "In some possible worlds, Mithridates wore a chocolate crown." But we can now see that the ordinary objects that figure in such sentences, like Mithridates, are modally thick: to be *Mithridates* is for indefinitely many counterfactuals to be true of one. If possible worlds are modally extensionless, then not even thin modal claims about ordinary objects can be paraphrased into the possible-worlds vocabulary without change of content.

had a clearer way of talking through one's modal claims. And so we find philosopher after philosopher observing how hard it would be to do your thinking about modal issues without possible worlds.

There were two pernicious consequences. The first was that a great many arguments came to assume the equivalence of claims in the two vocabularies; those arguments must now be reassessed, because possible worlds *don't* provide a clean way of keeping track of our ordinary modal claims: as we've just argued, possible-worlds renderings of ordinary counterfactuals don't preserve the contents of those counterfactuals. The second was that philosophers stopped thinking hard enough about modality, because they took it that it was something they *already* (basically) understood.

It should be obvious that to acquire a funny accent is not to understand anything you didn't understand before. But we now have an argument to add to that truism. Possible worlds are not a transparent alternative representation for our ordinary modal assertions, and an aid to understanding. And that's a problem not just for Lewis, but for everyone who talks that way, which means, most analytic philosophers. The peculiar dialect *doesn't* capture, and is not a passable surrogate for, the content of ordinary modal discourse. It's a mistake to think that manipulating the pictures associated with the dialect is going to help you understand the mysteries of modality.

References

Ainslie, G., 1992. *Picoeconomics*. Cambridge University Press, Cambridge.

Bremer, M., 2003. Is there an analytic limit of genuine modal realism? *Mind*, *112(445)*, 79–82.

Chihara, C., 1998. *The Worlds of Possibility*. Oxford University Press, Oxford.

Chisholm, R., 1948. The problem of empiricism. *Journal of Philosophy*, *45(19)*, 512–517.

Divers, J., 1997. The analysis of possibility and the possibility of analysis. *Proceedings of the Aristotelian Society*, *97*, 141–161.

Divers, J., 1999. A genuine realist theory of advanced modalizing. *Mind*, *108(430)*, 218–239.

Divers, J., 2002. *Possible Worlds*. Routledge, London.

Divers, J. and Melia, J., 2002. The analytic limit of genuine modal realism. *Mind*, *111(441)*, 15–36.

Divers, J. and Melia, J., 2003. Genuine modal realism limited. *Mind*, *112(445)*, 83–86.

Fara, M. and Williamson, T., 2005. Counterparts and actuality. *Mind*, *114(453)*, 2–30.

Fine, K., 1975. Critical notice: D. Lewis, *Counterfactuals*. *Mind*, *84(336)*, 451–458.

Forbes, G., 1985. *The Metaphysics of Modality*. Clarendon Press, Oxford.

Gentner, D. and Rattermann, M. J., 1991. Language and the career of similarity. In Gelman, S. A. and Byrnes, J. P., editors, *Perspectives on Language and Thought: Interrelations in Development*, pages 225–277, Cambridge University Press, Cambridge.

Gerber, S., Mayerik, V., Colan, G., Brunner, F., Buscema, J., and Infantino, C., 2008. *Howard The Duck Omnibus*. Marvel Comics, New York.

Housman, A. E., 1965. *Collected Poems*. Holt, Rinehart and Winston, New York.

Hughes, G. E. and Cresswell, M. J., 1996. *A New Introduction to Modal Logic*. Routledge, New York.

Kahneman, D., Slovic, P., and Tversky, A., 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press, Cambridge.

Kaplan, D., 1979. Transworld heir lines. In Loux, M., editor, *The Possible and the Actual*, pages 88–109, Cornell University Press, Ithaca.

Kazan, E., 1954. *On the Waterfront*. Horizon Pictures/Columbia Pictures, ???

Kripke, S., 1980. *Naming and Necessity*. Harvard University Press, Cambridge, Mass.

Lewis, C. I., 1946. *An Analysis of Knowledge and Valuation*. Open Court, La Salle.

Lewis, C. I., 1956. *Mind and the World-Order*. Dover, New York.

Lewis, D., 1973. *Counterfactuals*. Harvard University Press, Cambridge, Mass.

Lewis, D., 1983–1986. *Philosophical Papers*. Oxford University Press, Oxford.

Lewis, D., 1986. *On the Plurality of Worlds*. Blackwell, Oxford.

Lewis, D., 1999. *Papers in Metaphysics and Epistemology*. Cambridge University Press, Cambridge.

Lycan, W. and Shapiro, S., 1986. Actuality and essence. In French, P., Uehling, T., and Wettstein, H., editors, *Midwest Studies in Philosophy XI: Studies in Essentialism*, pages 343–377, University of Minnesota Press, Minneapolis.

MacBride, F., 2001. Can the property boom last? *Proceedings of the Aristotelian Society, 101(3)*, 225–246.

Mandler, M., 2001. A difficult choice in preference theory. In Millgram, E., editor, *Varieties of Practical Reasoning*, MIT Press, Cambridge, Mass.

Medin, D., Goldstone, R., and Gentner, D., 1993. Respects for similarity. *Psychological Review, 100(2)*, 254–278.

Melia, J., 1992. Against modalism. *Philosophical Studies, 68(1)*, 35–56.

Melia, J., 2003. *Modality*. McGill-Queen's University Press, Montreal and Kingston.

Millgram, E., 2005. *Ethics Done Right: Practical Reasoning as a Foundation for Moral Theory*. Cambridge University Press, Cambridge.

Millgram, E., 2009. *Hard Truths*. Wiley Blackwell, Oxford.

Mitchell, J., 1999. *Eccentric Lives and Peculiar Notions*. Black Dog and Leventhal Publishers, New York.

Moore, A., Gibbons, D., and Higgins, J., 2005. *Watchmen*. DC Comics, New York.

Nagel, T., 2007. The value of inviolability. In Bloomfield, P., editor, *Morality and Self-Interest*, pages 102–??, Oxford University Press, Oxford.

Paseau, A., 2006. Genuine modal realism and completeness. *Mind, 115(459)*, 721–729.

Plantiga, A., 1987. Two concepts of modality. *Philosophical Perspectives*, *1*, 189–231.

Preuss, A., 2007. Conjunctions, disjunctions and Lewisian semantics for counterfactuals. *Synthese*, *156*, 33–52.

Priest, G., 2001. *An Introduction to Non-Classical Logic*. Cambridge University Press, Cambridge.

Putnam, H., 1975a. Brains and behavior. In *Mind, Language and Reality*, pages 325–341, Cambridge University Press, Cambridge.

Putnam, H., 1975b. The meaning of 'meaning'. In *Mind, Language and Reality*, pages 215–271, Cambridge University Press, Cambridge.

Shafir, E., 1993. Choosing versus rejecting: Why some options are both better and worse than others. *Memory and Cognition*, *21(4)*, 546–556.

Shalkowski, S., 1994. The ontological ground of the alethic modality. *Philosophical Review*, *103(4)*, 669–688.

Shope, R., 1978. The conditional fallacy in contemporary philosophy. *The Journal of Philosophy*, *75(8)*, 124–135.

Sider, T., 2003. Reductive theories of modality. In Loux, M. and Zimmerman, D., editors, *The Oxford Handbook of Metaphysics*, pages 180–208, Oxford University Press, Oxford.

Skyrms, B., 1976. Possible worlds, physics and metaphysics. *Philosophical Studies*, *30(5)*, 323–332.

Thompson, M., 2008. *Life and Action*. Harvard University Press, Cambridge.

Wilson, M., 2006. *Wandering Significance*. Oxford University Press, Oxford.

Yagisawa, T., 1988. Beyond possible worlds. *Philosophical Studies*, *53(2)*, 175–204.