Metaphysically Reductive Causation

Metaphysically Reductive Causation

Ned Hall and L. A. Paul

There are, by now, many rival, sophisticated philosophical accounts of causation that qualify as 'metaphysically reductive'. A good thing: these collective efforts have vastly improved our understanding of causation over the last 30 years or so. They also put us in an excellent position to reflect on some central methodological questions: What exactly is the point of offering a metaphysical reduction of causation? What philosophical scruples ought to guide the pursuit of such a reduction? Finally, how should answers to these latter questions affect one's assessment of the main contemporary approaches? That's the stuff we'll be investigating in this essay.

§1 will lay out our presuppositions. §2 will review a sample of philosophical accounts. Then comes the main event: §3 will look in detail at the foregoing methodological questions, closing with a reconsideration of our sample accounts, in light of what we've found.

§1 Framework

§1.1 Metaphysical assumptions about causation

We will mostly assume that the fundamental causal relata are events,[1] These events are particulars, located in spacetime. We will not treat causation as a relation between *types* of events, although we grant that general causal claims can be made *apparently* concerning event types, viz. "overexposure to the sun causes sunburn."[2] We focus instead on singular causal claims – e.g., "Suzy's overexposure to the sun caused her sunburn."

---

[1] In fact we are both skeptical about this assumption. We assume it merely for simplicity and uniformity, as the vast majority of analyses treat causation as a relation between events. We will flag specific reasons for skepticism as we go along.

[2] One of us (NH) thinks such claims are really certain kinds of generalizations concerning token-level event causation. To think of them as expressing a metaphysically interesting relation between event types would be just as confused as thinking that "hens lay eggs" expresses some biologically interesting relation between a type of animal and a type of physiological product. In other words, the generic form of these claims easily misleads. See Nickel (2008) for helpful instruction on how not to be misled.

We also assume a broadly *reductionist* outlook, according to which facts about which events cause which other events are fixed, somehow, by (i) the facts about what happens, together with (ii) the facts about the fundamental laws that govern what happens. Minimally, that's a supervenience thesis: no two possible worlds differ with respect to what causes what without differing with respect either to what happens, or to what the fundamental laws are that govern what happens. But as will become apparent, the most important philosophical approaches to causation aim for something arguably stronger, namely, an account of causation that lays bare how causal facts are *grounded in* or *depend upon* these more basic facts. We'll come back to this point as we proceed, and will introduce some of the most influential accounts in §2, below.

As to (i), we take these to include all facts about what particulars exist at what times, and what categorical properties and relations these particulars instantiate. (Perhaps we can be bolder still, and take (i) to be exhausted by the facts about the total history of complete physical states that the world occupies; we won't need to take a stand on this.) As to (ii), the laws we have in mind are the fundamental dynamical laws of the sort that, we can hope, current physics is in the process of revealing to us. (These are to be distinguished from the so-called "laws" of the special sciences.) Taking such physics as our model, we think of these laws as something like rules that determine how complete physical states of the world generate successive physical states (see Maudlin 2007b).

We shall not investigate the metaphysical nature of these laws, although we *will* assume, purely for the sake of simplicity, that they are deterministic, and that they permit neither backwards causation nor causation across a temporal gap.[3] Perhaps they somehow consist in mere regularities (Loewer 1996). Perhaps they rest on firmer metaphysical foundations – e.g., necessary connections between universals (Armstrong 1983). Perhaps they are metaphysically primitive (Maudlin 2007b). Any account of laws that does not build on an antecedent notion of cause can serve as background for the sorts of issues confronting reductive analyses that we will be considering.

---

[3] Despite appearances, these constraints on the laws do not require causal notions for their articulation. Determinism is just the thesis that two nomologically possible worlds that agree on their histories up to time t agree simpliciter. The prohibition on backwards causation can be understood as the requirement that spacetime contain no closed time-like curves. And the prohibition on causation at a temporal distance can be implemented by requiring that, for any two nomologically possible worlds w_1 and w_2, if the complete physical state of w_1 at t_1 is the same as a complete physical state of w_2 at t_2, then the t_1-probability distribution over possible futures at w_1 is identical to the t_2-probability distribution over possible futures at w_2 (in other words, the present state of the world renders facts about the past irrelevant to what happens in the future).

§1.2 Neuron diagrams

In laying out examples, we make extensive use of "neuron diagrams" (popularized by Lewis: see in particular his 1986b). Here is a sample:

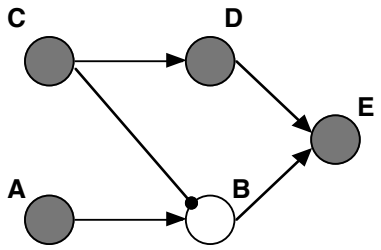

Figure 1

Circles represent neurons; arrows represent stimulatory connections between neurons; lines ending with black dots represent inhibitory connections. Shading a circle indicates that the neuron fires, with the temporal order read from left to right. Bold capitals name neurons, italicized capitals events of their firing. Thus, in Figure 1, neurons **A** and **C** fire simultaneously; **C** sends a stimulatory signal to **D**, causing it to fire, while **A** sends a stimulatory signal to **B**. But, since **C** also sends an inhibitory signal to **B**, **B** does not fire. Finally, **D** sends a stimulatory signal to **E**, causing it to fire. Figure 2 shows what would have happened if **C** had not fired:

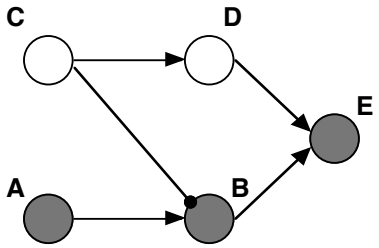

Figure 2

Neuron diagrams earn their keep by representing complex situations clearly and forcefully, allowing the reader to take in at a glance their central causal characteristics. However, they can also mislead: used carelessly, they can oversimplify an example, draw unwarranted attention to certain features of a case and underemphasize others, and even outright misrepresent the causal structure of the case. In addition, there are interesting problems, some of which we will mention below, that their use may obscure from view. So their prominence in this essay should not suggest that we think that every interesting and important feature of an example can be boiled down

to a neuron diagram, and where necessary we will take pains to highlight those features that cannot.

§1.3 A subtlety about metaphysics and reduction

We are interested in focusing on philosophical treatments of causation that see causal relations among events as somehow *metaphysically dependent* upon more metaphysically basic facts (concerning what happens, and what the fundamental laws are). That is, we're interested in *reductive* accounts of causation. That can raise a puzzle, though, concerning what the point of disagreement could be between rival such accounts. This puzzle is quite general, and has nothing per se to do with causation. All the same, it is important to get it out on the table, since attention to the philosophical issues it raises will frequently matter in what follows.

Here it is, in the (very!) abstract. Billy and Suzy, let us suppose, are having a dispute about the nature of some philosophically interesting category X. (X might be free will, or identity through time, or the nature of moral facts, or causation….) They have what is, apparently, a *factual* disagreement about X: Billy says it's one thing, Suzy says it's another. But let us further suppose that they agree *completely* on what belongs to the fundamental structure of reality: say, they both think that reality consists in a succession of complete physical states in space and time, which succession is governed by certain fundamental laws. And just to be clear, they have no disagreements about what the structure of these states might be, or what space and time are, or what the metaphysical nature of fundamental laws is, etc. (Maybe they both think it's all fundamentally atoms in the void, moving about under the direction of Newtonian laws.) And yet they disagree about the nature of X.

The puzzle is that it can seem that there is no longer anything *substantive* for this agreement to be *about*. "You both agree on the fundamental facts," one wants to say. "Isn't the rest just talk?" Compare an argument Billy and Suzy might have about whether *viruses are alive*. Neither is a vitalist: both agree that facts about what's alive and what's not somehow reduce to biochemical facts – about which, in turn, they have *no* disagreement. How could their dispute about viruses possibly be genuinely substantive or factual?

Maybe it couldn't. That is certainly *one* intelligible, defensible view. Quite generally, one might adopt the following meta-metaphysical principle: *Any metaphysical dispute over the nature of some feature of reality X must, in order to be genuinely substantive and non-terminological, trace to a dispute over the nature of fundamental reality*. Maybe X itself is recognized to be a feature of fundamental reality, in which case the principle obviously holds. Maybe X is some not-very-philosophically-interesting feature – for example, the dispute might concern whether it is raining outside. Again the principle holds,

since we can reconstruct this dispute, in a rather labored fashion, as a dispute about whether the detailed disposition of the fundamental facts is such as to make it the case that it's raining outside. But maybe X is some vastly philosophically significant feature of reality, nevertheless recognized to be non-fundamental: and here, the thought goes, the principle yields something of value, by showing us that if we want our dispute about the nature of X to be *genuine*, then we had better figure out how it hinges on a disagreement about fundamental reality.

We don't know whether this principle is correct. All we wish to urge here is that however attractive and obvious it may seem when one looks at *some* debates – e.g., Billy and Suzy's debate about whether viruses are truly alive – one should not *automatically* assume that it holds across the board. Suppose, this time, that Billy and Suzy are arguing over whether the statue is identical to the lump of clay that it is made from.[4] Once again, they agree that all that exists, fundamentally, are atoms in the void, subject to such-and-such fundamental laws. What they disagree about is *how many ways those atoms combine to compose nonfundamental objects*.[5] In cases such as this, we think it hasty to insist that their debate cannot possibly be substantive.

An apt rejoinder: "Fine. Maybe their debate *is* substantive, after all. *But how??* Explain, please." Here's why, in our view, the rejoinder is apt: even though (*we* think!) the meta-metaphysical principle cannot reasonably be assumed a priori, it *can* be taken to locate the burden of proof, in that one who rejects it, in a specific case, owes an account of exactly *how* a debate about non-fundamental feature X can be substantive, in the face of full agreement about what's fundamental.

Example. Return to Billy and Suzy's statue/clay debate. The meta-metaphysical argument that this debate is substantive, even though both parties agree that fundamentally, it's all just law-governed atoms in the void, might proceed as follows: Statues, and lumps of clay, are not themselves fundamental entities. Rather, they are *constituted* by such entities. But (and here comes the crucial move) there can be substantive disagreement about what is required for *genuine constitution*. More specifically, Billy and Suzy might *both* hold that the statue and lump of clay are constituted by particles, by being mereological fusions of particles. (On mereology – the theory of parts and wholes – see Simons 1987, Lewis 1991.) And they might agree that the particles that are ultimate parts of the statue are all and only the particles that are ultimate parts of the piece of clay. But they might *disagree* over a basic question of mereology: namely,

---

[4] See for example Fine (2003).

[5] It's important for the example that facts about composition don't themselves count as fundamental – plausible, since they concern a relation between fundamental entities and nonfundamental entities.

whether, given some entities (the particles, in this case), they have a *unique* mereological fusion. It is because *that* disagreement is substantive that their disagreement about whether the statue and the piece of clay are one and the same object is, too.

We recognize that those attracted to the meta-metaphysical principle will grumble. Maybe they will want to say that it's just a confusion to think that the correct principle about mereological fusions is at all controversial. Whatever. Here, we wish only to caution you, the reader, that it is philosophically naïve to treat as obvious that agreement over fundamental reality must render any disagreement over the nature of causation purely terminological.

Finally, suppose some such disagreement *is* terminological: really, it concerns not the *facts* about what causation is, but rather which of the many aspects of reality that both sides agree there are to call "causation". For all that, it may be a disagreement well worth having – and so not at all "terminological", in the dismissive sense that might apply to saying that debate was "merely verbal" (compare a debate over whether whales are fish; see Chalmers 2011). In general, it matters for our intellectual aims – especially our explanatory aims – that we categorize things *well*. And so it might matter quite a lot what we choose to call "causation". The dismisser says, "You agree on what's fundamental; the rest is just talk." To which a good reply is, "Yes – but it can sometimes matter quite a lot that we construct our talk *well*."

§2 The map of rivals

This section sketches some of the most significant rival approaches to providing a philosophical account of causation. We start with an approach that has unjustly fallen into disfavor.

§2.1 Regularity accounts

What have been called "regularity" accounts of causation have been guided by two quite distinct ideas. First idea: Causal relations between events should be analyzed as *instances of lawful regularities*. Thus Davidson (1967) suggests, roughly, that when *C* causes *E*, there must be some suitable descriptions of these events – as, say, the F-event and the G-event, respectively – such that there is a law connecting F-events with G-events. (Davidson's own candidate for such a law unhelpfully includes the word "cause" in its statement: not a good idea, given that the paradigm examples of fundamental laws provided by physics never do so.) Second idea: What is distinctive of the causes of some event is that they lawfully *suffice* for it, at least in the circumstances (and: given determinism). Probably the best known account along these lines comes from Mackie (1965), although we will present it here in a form borrowed from Lewis (1973a): *C* causes *E* iff *C* and *E* both occur, and there is some suitably

chosen auxiliary proposition F describing the circumstances of *C*'s occurrence such that (i) in any nomologically possible world in which *C* occurs and F is true, *E* occurs; (ii) in some nomologically possible world in which F is true, and *C* does not occur, *E* does not occur. In short, *C* is an essential part of some set of conditions that is lawfully sufficient for *E*.

The two guiding ideas should be kept carefully separate. The first has insuperable problems, which we will return to below (§3.3.1). The second has more merit. But it also needs a more careful expression than that given above, since it is too unclear how to pick out the auxiliary proposition F. Here is a better way to proceed: Start with the observation that F must, presumably, include a description of the *other* causes with which *C* combines to bring about *E*. That leads to the suggestion that what is key is that the set S of causes of an event *E* should *collectively suffice* for that event, but should do so *non-redundantly*: i.e., no proper subset of S should suffice for *E*. Then S had better not include *all* the causes of *E*, occurring at *any* time, since later ones will render earlier ones redundant, and vice versa. So let us require merely that the set of causes of *E* that *occur at some given time* (before *E* occurs) non-redundantly suffice for *E*. That amendment still clearly does justice to the guiding idea. We arrive at the following provisional analysis: *C* is a cause of *E* just in case for some time t earlier than *E*, *C* belongs to a set of events occurring at t that non-redundantly suffices for *E*.

What remains is to say what "suffice" means. Here is a first pass. A set S of events suffices for (later) event *E* just in case the occurrence of those events lawfully guarantees that *E* occurs: in any nomologically possible world in which all the members of S occur, *E* occurs. Notice that on this reading, our regularity account does not in any interesting sense view causal relations as "instances" of "covering laws", in the way that Davidson's account did; rather, all that is required of the laws is that they draw a distinction between the nomologically possible and impossible. For that task, the laws of fundamental physics will do just fine.

Still, this account of sufficiency overlooks an important issue, since it will in general be possible for the events in S to occur jointly with other "inhibiting" events that act so as to *prevent* the occurrence of *E*. In figure 3, for example, it obviously doesn't follow from the fact that **C** fires, together with the "neuron laws," that **E** will fire (for what if **A** had fired?)[6]:

---

[6] We have often encountered, both in conversation and in print, the view that determinism entails (or even just *is*) the thesis that the causes of any event guarantee that event's occurrence. Not so: for even under determinism, the causes of some event do not guarantee that nothing occurs that could *prevent* those causes from bringing about that event.

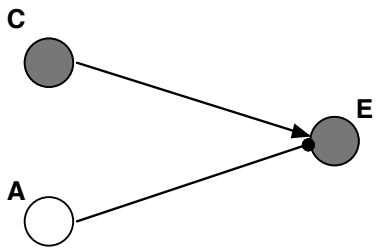

Figure 3

A better idea is to say that S suffices for *E* just in case, were the events in S to occur *without any interference*, *E* would occur. If we agree that such interference would require the occurrence of at least one other, contemporaneous event, then we can simplify, as well as remove any residual taint of nonreductivity: A set S of events occurring at some time t suffices for (later) event *E* iff, were the events in S the *only* events occurring at t, *E* would (still) occur. Calling a set *minimally* sufficient just in case it is sufficient, but no proper subset is, we thus arrive at the following *updated regularity account*: *C* is a cause of *E* iff *C* belongs to a set of contemporaneous events that is minimally sufficient for *E*. In the simple form displayed here, it is an attractive and useful example of the type.

§2.2 Counterfactual accounts

Counterfactual accounts of causation begin with the idea that, when *E* counterfactually depends on *C* (or for short, just "depends") – when, that is, it is the case that if *C* had not occurred, *E* would not have occurred – then *C* must be a cause of *E*. Promoted to a sufficient *and necessary* condition, that won't do: it is easy enough to have circumstances in which *C* causes *E*, even though backup processes would have brought about *E* in *C*'s absence (as figure 1 already shows). But as a sufficient condition on causation, it has struck many philosophers as exactly right – and therefore as an excellent starting point for a full-blown analysis of causation. Scan the literature on causation, and you will find a profusion of such analyses, departing in myriad different directions from this leading idea. We will by no means try to provide a comprehensive survey, but will sketch three especially interesting avenues.

The alert reader may be wondering how an account of *counterfactuals* can be developed that does not in some way rely on *causal* notions. An excellent question, which (along with related questions) we defer until §§3.3.2 and 3.3.3.

§2.2.1 Chains of dependence

A well-known and elegant approach comes from Lewis (1973a), who analyzes causation as the ancestral of counterfactual dependence: *C* causes *E* just in case there is a (possibly empty) set of events $\{D_1, D_2, ..., D_n\}$ such that D_1 depends on *C*, D_2 depends on D_1, ..., and *E* depends on D_n.

Figures 1 and 2 display a natural motivation for this approach. For it is clearly the case that *C* causes *E*; yet *E* does not depend on *C*. So it won't do simply to identify causation with counterfactual dependence. On the other hand, *D* clearly depends on *C*, and – *provided* we understand the counterfactual in a certain, specific way – *E* likewise depends on *D*. What way is that? As *non-backtracking*. We'll look in more detail at what this amounts to in §3.3.2, but for now we'll make do with the following idea: In constructing the counterfactual situation in which **D** does not fire (at the given time t), we *hold fixed* the state of **D**'s surroundings at t. So we hold it fixed that, at t, **B** likewise fails to fire. We do *not* to reason that if **D** had not fired, that would have to have been because **C** did not fire, whence **B** *would* have fired (hence, so too, **E**).

Notice that by taking the ancestral, the chains of dependence approach analytically guarantees that causation is transitive.

§2.2.2 Influence

The second account comes from Lewis's more recent work (2000, 2004a). In it, he replaces the simple relation of dependence with a more complicated relation of *counterfactual covariation*. Very roughly, *E* counterfactually covaries with *C* just in case (and to the extent that) variation in the manner of *C*'s occurrence would be followed by corresponding variation in the manner of *E*'s occurrence. The situation in which *C*'s absence would be followed by *E*'s absence is, Lewis thinks, a kind of limiting case. Following Lewis, say that *C influences E* just in case *E* counterfactually covaries with *C* to a sufficient extent. (Also following Lewis, we will leave it vague what counts as "sufficient".) Lewis's proposal is that causation is the ancestral of influence.

§2.2.3 De facto dependence

The third approach, championed by Yablo (2004) and by some advocates of "structural equations" (of which more in a moment), identifies causation with what Yablo has called "de facto dependence": *E* de facto depends on *C* just in case had *C* not occurred, and had other suitably chosen factors *temporally between C* and *E* been held fixed, then *E* would not have occurred. The trick is to say what "suitably chosen" means, and to give clear and systematic truth conditions for this more complex kind of counterfactual. We'll bypass the second of these issues, elaborating briefly on the first by considering one simple example of a de facto dependence account.

We draw our example from the *structural equations* literature, partly because structural equations approaches to causation currently enjoy a fair bit of popularity (see for example Woodward 2005 and Pearl 2000). To lay out the example we'll need to digress, in order to explain what is distinctive about the structural equations approach. Here's the idea: In order to analyze the causal structure of any situation, we must first provide a "causal model" for it. The elements of this model consist of (i) *variables*, (ii) a range of possible *values* for each of the variables, (iii) a specification, for each variable, of which other variables (if any) it *immediately functionally depends on*, and (iv) "structural equations" that describe this dependence. Thus, if the situation we are modeling is one in which Suzy throws a rock at a bottle, breaking it, we might construct a simple causal model by assigning a variable to the bottle whose values represent its different possible states (e.g. broken, fractured, unharmed), and assigning a second variable to Suzy's throw whose values represent the strength of the throw (and whether it happens at all). Our model will represent the first variable as functionally depending on the second, according to an equation that says, in effect, that the bottle will break (the bottle-variable will take the value 'broken') iff Suzy throws with a strength above a certain threshold.

It is an excellent question, inadequately addressed in the literature, precisely what principles should guide the construction of a causal model. One could be forgiven for suspecting that these principles really require one to figure out what causes what, in the situation to be modeled, and then to select variables and functional relationships among them to fit. Never mind. (Though see Hall 2007 for discussion.) Given that we are going to be confining ourselves mostly to examples represented by neuron diagrams, it will in general be obvious how to construct an appropriate causal model: First, assign a variable to each neuron, which can take on a range of values corresponding to each different way that that neuron can fire, reserving one more value for the situation in which it does not fire at all. Next, stipulate that each such variable immediately functionally depends on the variables for those neurons that have a direct "incoming" connection to it, either stimulatory or inhibitory. And finally, write the functional equations down so as to capture exactly how the various possible firing patterns for the input neurons to a given neuron will determine whether and how it fires.

Here is the crucial innovation: with causal model in hand, you are in a position to give systematic truth-conditions for a *novel sort of counterfactual*: namely, one whose antecedent specifies values for *arbitrarily many* variables. And this expanded set of counterfactuals provides the tools for correspondingly novel analyses of causation.

Suppose for example that we have a causal model for the events depicted in Figure 1.

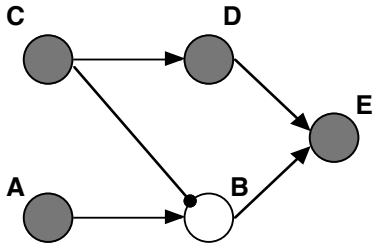

Figure 1

It's pretty clear how to use the model to construct a counterfactual situation in which **C** does not fire: Take the actual value of the C-variable (1, for firing) and change it to 0 (non-firing). Make appropriate adjustments downstream, recalculating the values for every variable that depends, either immediately or remotely, on this variable. In doing so, *do not* change the actual value of the A-variable. Result: a situation in which the counterfactual value of the D-variable is 0, of the B-variable 1, and of E-variable 1. In words: if **C** had not fired, then **D** would not have fired, but **B** *would* have fired, and therefore so would **E**. So far, so good. (And so far, nothing new.)

Now suppose we want to consider a situation in which **C** does not fire, but **B** *also* fails to fire (never mind *why*). When, as here, the antecedent stipulates the value for some "endogenous" variable (i.e., a variable whose value functionally depends on other variables explicitly represented in the model), then in constructing the counterfactual situation we simply *ignore* those functional equations that would otherwise have fixed the value of this variable. (It helps to imagine that the endogenous variable gets tweaked by an outside *intervention*, that breaks that variable's connection to its input variables – hence the common label "interventionist" for accounts of causation like the one we are considering.) Thus, we set the value of the C-variable to 0, of the A-variable to 1 (its actual value), and the B-variable to 0. We then calculate the values for the D- and E-variables according to the appropriate functional equations, with the result that the D-variable has the value 0 and the E-variable also has the value 0. In words: if **C** had not fired, and **B** had (still) not fired, then **E** would not have fired.

Observe that this counterfactual allows us to say that *in a sense*, *E* in Figure 1 *does* depend on *C*; for *in fact* **B** does not fire (the B-variable has the value 0), and if we *hold this fact fixed*, then *E* depends on *C*. More generally, it is by means of such counterfactuals that those who pursue a "structural equations" version of a de facto dependence account of causation aim to analyze that relation.

Here, finally, is one specific proposal, drawn from Hitchcock (2001): Suppose that we have two events, *C* and *E*, and associated variables C and E. And suppose

that we have some appropriate causal model of the situation in which *C* and *E* occur. Say that there is a "path" from *C* to *E* just in case there is a possibly empty set of variables $\{D_1, D_2, \ldots, D_n\}$ such that D_1 immediately functionally depends on C (and possibly other variables; we omit this qualification henceforth), D_2 immediately functionally depends on D_1, … and E immediately functionally depends on D_n. Now (departing slightly from Hitchcock for ease of exposition) suppose that there are one or more variables that are *not* on this path, such that if we hold them fixed at their actual values, then *E* depends on *C*. More exactly, the counterfactual circumstance we represent by setting the C-variable to 0, and holding those other off-path variables fixed at their actual values, is one in which the E-variable also gets set to 0. Then, adopting Hitchcock's terminology, we can say that the given path from *C* to *E* is an "active route". A simple proposal results: *C* is a cause of *E* iff there is an active route from *C* to *E*. For example, in figure 1 the *C-D-E* route is active, as witness the fact that if **C** had not fired and **B** had also not fired, then **E** would not have fired. By contrast, there appears to be no active route from *A* to *E*: for the only candidate is the *A-B-E* route, and holding fixed any combination of the C- and D-variables fails to make it the case that *E* depends on *A*.

This is *not* the only way to construct a de facto dependence approach, or even a structural equations variant thereof. But it will provide an attractively simple illustration of the approach in the pages ahead; observe in this regard that it is crystal clear what constrains the choice of factors to be "held fixed" (at least, *modulo* the provision of an appropriate causal model). The reader is invited to contrast Yablo's (2004) discussion of this matter, which is far more intricate.

§2.3 Probabilistic accounts

Probabilistic accounts of causation are closely related to counterfactual accounts, although more naturally suited to treating causation in the indeterministic domain. Consider what each account takes as the central feature of the causal relation: For counterfactual accounts, it is that the effect counterfactually depends on the cause; for probabilistic accounts, it is that the effect *probabilistically* depends on the cause – that is, the probability that the effect occurs, *given* that the cause occurs, is higher than the probability that the effect occurs, given that the cause does not occur. (For sophisticated examples of probabilistic accounts, see Eells 1991, Kvart 2004, and Ramachandran 2004.)

We will have no more to say about probabilistic accounts in this essay. Not because we consider them unimportant. Rather, we overlook them in part because our focus is on causation in the deterministic domain – a domain in which all probabilities are one or zero, making probabilistic relations too crude an instrument for un-

derstanding causation[7] – and in part because the relations between probabilistic and counterfactual accounts are *so* close that problems for one often carry over to the other.[8] We will give just one illustrative example. In Figure 1, it is obvious that *E* does not depend counterfactually on *C*, since if *C* had not occurred, the backup process initiated by *A* would have brought about *E*. But for exactly the same reason, *E* does not depend probabilistically on *C*: the probability that *E* occurs is independent of whether *C* occurs (understanding the example now to involve appropriately "chancy" neurons). So each account will have to adopt some strategy for circumventing this and other kinds of examples. And what one finds when one surveys the problem cases is that the available strategies are remarkably similar.

§2.4 Transference accounts

The recent literature has seen some interest in accounts of causation according to which it essentially involves the *transfer* of some sort of quantity from cause to effect. Typical accounts turn to physics in search of the right quantity. For example, Fair (1979) takes it to be energy, while Dowe (2000) and Salmon (1994) allow it to be any sort of quantity that is, according to the fundamental physical laws, conserved. Other "transference" accounts (as we will call them) are more metaphysical: Ehring (1997), for example, takes causation to consist (at least in part) in the transfer of tropes, i.e. particularized properties.

We very much doubt that *pure* versions of such accounts – ones that contain no admixture of ideas borrowed from regularity or counterfactual approaches – have a prayer of working: for reasons we'll review in §3.3.5, it simply won't fly to *identify* causation with the transfer of some special quantity. Still, it is possible that a fully adequate account of causation should incorporate elements of some transference account. For example, perhaps the best way to deal with cases of preemption such as Figure 1 is first to discern the patterns of transfer of the relevant quantity or quantities, and then to look at more abstract relations of counterfactual dependence or sufficiency, etc., that these transfers exhibit. Transference accounts also prove very useful as a foil for drawing out a variety of issues having to do with causal relations that essentially involve omissions (see chapter 4 of Hall and Paul 2011).

---

[7] This assumes, as is surely appropriate, that the probabilities are understood as *objective chances,* and not identified with or constructed out of *subjective credence.*

[8] Having said that, causation in the probabilistic domain raises several fascinating puzzles that have no obvious analogues in the deterministic domain. These are not puzzles about probabilistic accounts *per se*; but they are interesting enough to deserve close study. See for example Frick 2009, or Hall and Paul 2011, ch. 2.

§3 Methodology

With sketches of some of the most important accounts of causation in hand, it's time to jump into our methodological questions. If all goes well, we'll emerge with a cleaner understanding of what an account of causation ought to be aiming to accomplish, and, consequently, a better appreciation of the point of the close, examples-based kind of analysis that is so common in the literature.

§3.1 Varieties of analysis

Suppose a philosopher offers up, as part of some philosophical theory, some biconditional[9] "A iff B". Of course, in typical cases she will really be presenting a *schema*. It might have the form "C is a cause of E iff – ", or "S knows that p iff – ", of "F is an intrinsic property iff – ", etc., with the blank filled in in some interesting way; we're all familiar with many examples of the type. What exactly do we expect from such a schema? What are the standards of success that our philosopher is trying to meet?

Well, every instance of the schema ought to turn out to be *true*. But that's not very helpful, for more or less obvious reasons. The instances of the biconditional might be true but uninformative ("C is a cause of E iff C is a cause of E"), or might merely *happen* to be true (as opposed to being true a priori, or true with some kind of analytical or metaphysical necessity).[10] Even if we are content to say that we expect instances of the biconditional to be informative (without saying precisely what this means), and that we expect them to hold with some kind of necessity (without saying precisely which kind), we are missing further useful distinctions. Here they are.

§3.1.1 Mere necessary connection

First, our philosopher might *merely* be aiming to highlight an interesting, and in some sense "necessary", connection between the two sides of her biconditional – without claiming, further, that either side can in any sense be "explained away" in terms of the other. Example: many philosophers find it plausible that property F is *intrinsic* just in case, for any two possible objects x and y that are *perfect duplicates* of each other, either both have F or neither has F. Someone might offer this biconditional as a moderately informative, useful connection between intrinsicness, duplication, and modality – without any aspiration to turn it into a reductive analysis or definition of "intrinsic property".

---

[9] Or maybe, less ambitiously, just a *conditional*, in one direction or the other.

[10] Though one advantage of working with schemata is that it's hard to produce one where all of its instances just *happen* to be true.

§3.1.2 Stipulative definition

Next, our philosopher might be presenting a *stipulative definition*. In some contexts, this move is just fine – e.g., when introducing explicitly technical vocabulary, and explaining how it is to be understood. But in other contexts it's not so fine, especially if the move is designed merely to allow one to avoid seriously grappling with counterexamples. Suppose, for instance, that you have become deeply enamored of a simple counterfactual analysis of causation: *C* is a cause of *E* iff *C* and *E* both occur, and if *C* had not occurred, *E* would not have occurred. You're perfectly aware of cases of preemption that seem to spell doom for this analysis. (Figure 1 will do.) But rather than tinker with your analysis, you decide to offer it up as a stipulative definition of what you shall henceforth mean by "cause". Then granted, no one can complain that preemption presents you with a *counterexample*.[11] But all you've done is force the complaint to be registered in a different mode: Now the worry will be that you have drawn a *useless* distinction (not to mention that you have drawn it using familiar words in misleading ways) – or, at any rate, that you have *overlooked* a valuable distinction (viz., the distinction *we* mean to be drawing by our use of the word "cause").

Having said all this, it can sometimes be a valuable exercise to ask (non-rhetorically!), of an analysis that runs afoul of some example, "What would be wrong with avoiding the counterexample, simply by treating this analysis as a stipulative definition?" We'll come back to this point below.

§3.1.3 Conceptual analysis

Next, our philosopher might offering up a good, old-fashioned *conceptual analysis*. We doubt that there is any clear and widely-agreed upon conception of just what conceptual analysis is, or what its standards of success are. We'll distinguish two options. First, you might have a Fregean view of concepts, according to which they are a kind of abstract object which the mind *grasps* in having thoughts. Maybe some of these concepts are *structured*, in ways that involve other concepts as *constituents*. Then conceptual analysis could aim to put on display the way in which one concept is constructed out of other, more basic concepts.

We're going to set this idea aside, as its philosophical presuppositions strike us as too implausible. There is an alternative, which is to treat conceptual analysis as a kind of *project in empirical psychology*. Start with the view that concepts are psychological

---

[11] Compare Goodman's (fictional!) "proof that p": "Zabludowski has insinuated that my thesis that p is false, on the basis of alleged counterexamples. But these so-called 'counterexamples' depend on construing my thesis that p in a way that it was obviously not intended – for I intended my thesis to have no counterexamples. Therefore p."

structures realized in the brain that enable specific kinds of psychological activity. (E.g., an agent who possesses the concept *cat* is thereby able to have thoughts about cats. To possess the concept *cat* just is to have realized in one's brain a certain kind of structure with certain distinctive functional relationships to the rest of one's psychological economy, and to the outside world.) Then an *analysis* of a concept can be thought of as a *theory of how that concept functions* in actual human psychological economies.

If a philosopher presents herself as engaged in "conceptual analysis", it's a very good idea to ask her which of these conceptions of conceptual analysis she has in mind, if either. As noted, we don't find the first, Fregean conception terribly useful. The second is another story: there is *plenty* of worthwhile investigation to be done into how humans actually engage in causal reasoning. (For a small sampling of recent psychological literature, see Gopnik et al. 2004, Lombrozo 2010, Sloman 2005, Wolff 2007.) But there is also a glaring question as to what exactly armchair philosophical speculation has to contribute to such an investigation. A reasonable answer (one that many psychologists themselves would happily accept): Armchair speculation can, done with sufficient care and creativity, generate hypotheses worth testing against empirical psychological data.[12] Still, it's safe to say that the more success psychology achieves at uncovering the structure of such reasoning, the less such work there will be for philosophers.

There is another role for empirical psychology in a very different kind of philosophical project; one involving ontological reduction. We'll come back to it below.

§3.1.4 Ontological reduction

The fourth kind of analysis our philosopher might be presenting is an *ontological reduction* of causation. That is, she might be intending that one side of her biconditional puts on display how facts about what causes what *reduce to* ontologically more basic facts. We'll spend the most time exploring this option.

As a helpful illustration, forget about causation for the moment, and focus on *the direction of time*. Suppose we take as ontologically fundamental relations of *temporal betweenness*. That is, we are not looking to analyze "time t_1 is between time t_2 and time t_3". But we *are* looking to analyze "time t_1 is *earlier* than time t_2". What's more, we seek analysis-as-ontological reduction, in that we think that *what it is* for one time to be earlier than another can be explained in more ontologically basic terms. Then here is an ontological reduction that many have found attractive (see for example Albert

[12] See for example Lombrozo 2010.

2000): First, we hold that our universe has one *low-entropy temporal end*.[13] Then we hold that, in the temporal direction that proceeds away from this end, global entropy always increases.[14] And now we can say: time t_1 is earlier than time t_2 iff t_2 lies, relative to t_1, in the direction of global entropy increase.

There are a few things to notice about this example.

First, our biconditional doesn't carry reductive intent on its face, even if we add "it is necessary that" to the front of it. Suppose you think that it's just a primitive metaphysical fact what the direction of time is (cf. Maudlin 2007c). You might agree to the biconditional all the same – you will just take it to state a substantive fact about *entropy* (namely, that it *globally increases toward the future*). You might even think this fact is related to other facts so deep that, while it is strictly speaking nomologically possible for global entropy to *decrease* toward the future, this is so objectively unlikely as to be deemed impossible, for all practical purposes. So if, by contrast, you view the biconditional as laying out how facts about the past/future direction are reducible to other, more basic physical facts, then you should just *say so explicitly* – and not try to pretend that your intent can be adequately captured by insisting that the biconditional holds with some kind of necessity.[15]

Second, it's actually not so clear *what* kind of necessity should attach to this biconditional. To be sure, you *might* claim that, as a matter of *metaphysical necessity*, one time is earlier than another iff it lies in the direction of lower global entropy. But you don't *have* to claim any such thing. You might insist only that the biconditional holds in all possible worlds with laws of nature that allow for facts about entropy. Or you might be suspicious of the very notion of metaphysical possibility, as distinct from nomological possibility, and be willing only to say that it *could have turned out* (but didn't) that the past/future asymmetry was not grounded in the direction of global entropy increase.

The issues here – about how exactly to understand metaphysical necessity, and its relation to ontological reduction – are subtle, and obviously we don't pretend to have settled them. But we insist – and this, really, is the important point – that it is *not* an effective dialectical maneuver against a proposed ontological reduction merely to devise a conceivable scenario that violates it. Imagine the following conversation:

---

[13] Equipped with a notion of temporal betweenness, we can easily say what it is for one time to be a *temporal end*: it is not between any other two times.

[14] Or: sometimes increases and never decreases. Or maybe we get fancier still, and allow for very short-lived, occasional decreases, so long as the general trend is toward increasing.

[15] Which is not to say that it *doesn't*. It's just that it seems to us more accurate to say that you take the biconditional to be necessary *because* you take it to describe a relation of ontological reduction, and not that you take it to describe a relation of ontological reduction because you take it to be necessary.

Suzy: I think that the direction of time reduces to the direction of entropy increase; *what it is* for one time to precede another is for the second time to reside, relative to the first, in the direction of entropy increase.

Billy: But that can't be. For it is surely conceivable that entropy *decreases* toward the future. And what is conceivable is metaphysically possible. And if you are right that temporal direction is reducible to the direction of entropy increase, then this must be so *necessarily* – which it is not. So your view stands refuted.

Suzy: No, it doesn't.

(Pause.)

We side with Suzy.[16] In fact, this is one of those cases where dodging an apparent counterexample by means of stipulative definition can be a very helpful tactic. That is, what Suzy should go on to say is this:

Suzy: "I'm not giving an account of what *you* mean by 'the direction of time'. But I stipulate that this is what *I* shall mean by that expression so that I can go on to develop an account of the metaphysical nature of the direction of time. And I now challenge you to show why, by doing so, I'm making any sort of serious mistake, or missing something of importance."

If, in reply, the only things Billy can point to are off-the-cuff intuitions about outré cases, then Suzy should remain unimpressed.

Third, there is really no hope of viewing this reduction of facts about past and future to facts about entropy as *conceptual analysis* – at least, not of the second, psychological type, and plausibly not of the first, Fregean type either. It is striking that this fact does not impugn the philosophical interest of the analysis in the slightest.

Fourth, there may still be a role – albeit indirect – for empirical psychology in this sort of analysis. Our ordinary temporal concepts have a certain structure. It's the job of empirical psychology to articulate that structure. Not every aspect of that structure needs to be mirrored in, or even consistent with, Suzy's account of temporal direction. (Again, she's just not in the business of conceptual analysis.) *But:* It *should* be possible for her to explain, by means of her account, what sorts of objective temporal structures out there in the world our ordinary concepts are *answerable to*. Given how the world is, according to her, *actually* temporally structured, how is it that

[16] In saying this, we're obviously denying the conjunction of views that (i) conceivability is a guide to metaphysical possibility (contra for example Chalmers 2002), and (ii) reductive analyses of the kind Suzy is offering here hold must be metaphysically necessary, if true.

our ordinary ways of conceptualizing temporal structure work as well as they do? She ought, at least at the end of the day, to be able to say. And successful empirical psychological inquiry into the nature of our ordinary temporal concepts is essential, if we are to see what this explanatory demand placed upon her account amounts to.

Finally, the clarity and interest of the example ought to help allay fears about the philosophical coherence or legitimacy of talk of "ontological reduction". It's perfectly reasonable to wonder what exactly is going on when a philosopher announces that *what it is* for such-and-such a fact to obtain is for such-and-such other fact to obtain; or (equivalently, in our view) that this fact holds *in virtue of* that fact, etc. We'll happily go further: it's perfectly reasonable to be on one's guard against overly sloppy, cavalier, or mystifying appeals to such notions of ontological "grounding". But caution should not give way to wholesale rejection of the kind of metaphysical inquiry that seeks substantive, illuminating answers to questions of the form "What is it for such-and-such fact to obtain?" We have a good enough collective grip on the distinction between more and less ontologically fundamental facts to be able to evaluate proposed answers to such questions with respect to how substantive and illuminating they are. Part of the reason for highlighting philosophical positions such as the foregoing one about the direction of time is precisely to remind ourselves, by means of a vivid example, that we do indeed understand what is being asked by a question such as "What is it for one time to be earlier than another?" and can indeed recognize a substantive and illuminating answer when we see one. And we can recognize this despite the fact that we may have no explicit theory of the "in virtue of" relation.

Now for the punch line: In the case of causation, we propose that what is of primary interest is whether a philosophical account of causation, *understood as an ontological reduction*, can be given, and if so, what are the plausible forms it might take.

§3.1.5 Some bad habits to avoid

With this aim in mind, we can see the importance of guarding against two sorts of bad habits. The first is to make assumptions about a test case implicitly grounded in knowledge of the causal structure of that case. We'll give an extended example.

Late preemption is a particularly thorny kind of causal preemption. Suzy and Billy both throw rocks at a bottle, with perfect accuracy; but Suzy's rock gets there first, shattering it. If Suzy hadn't thrown, the bottle still would have shattered. So a simple counterfactual analysis of causation fails in this case, as does Lewis's "chains of dependence" account, and, arguably, his "influence" account. (See Hall and Paul 2011, ch. 3.) Now, Yablo, in arguing for the advantages of his de facto dependence account, claims that it can easily handle this sort of account: holding *fixed* the fact that Billy's rock *did not strike the bottle*, its breaking depends on Suzy's throw. But why be-

lieve that? Why assume, that is, that the counterfactual "If Suzy had not thrown, and if Billy's rock had (still) not struck the bottle, the bottle would not have broken" is *true*? Yablo offers no account of the truth-conditions of this counterfactual that yields this verdict; he simply takes its truth as intuitively obvious.

That's a mistake. To see why, try working out the truth-conditions: in the counterfactual situation that begins with Suzy not throwing, make some additional small change to the world so that Billy's rock fails to strike the bottle. (For more on the truth-conditions of counterfactuals, see §§3.3.2 and 3.3.4, below.) Alas, there are ever so many ways to do that – and some perfectly reasonable ways have the result that the bottle *does* shatter. Suppose, for example, that the bottle is perched on a post. Then one way it could come about that Suzy does not throw, that Billy does throw, and that Billy's rock somehow fails to strike the bottle, is for a gust of wind to knock the bottle off the post – in which case, fragile thing that it is, it shatters upon hitting the ground.

If we could make free use of *causal* facts about this case, then we could plausibly provide a successful recipe for constructing the relevant counterfactual situation, as follows: First, identify all the causes of the bottle-shattering; distinguish these from other factors that are non-causes. Next, construct a counterfactual situation in which Suzy does not throw, Billy throws, but in which other forces are introduced—let us provide them with the convenient label "God"—that cause the factors to be held fixed to obtain, *without in any way interfering with any of the processes that are, in actual fact, causally involved in the bottle-shattering*. So we allow God to do whatever it takes to Billy's rock to make it the case that it does not hit the bottle, as long as these interventions do not causally interact with the bottle itself.

We do not know why Yablo thinks it intuitively obvious that the counterfactual has the truth-conditions he needs it to have. But we speculate that it *seems* to only because he's holding the actual causal structure of the situation in the back of his mind, and letting this structure inform the way he understands the counterfactual (perhaps in the manner suggested in the last paragraph). No fair. Absent a proper *account* of the counterfactual's truth-conditions, he can't lean on a mere intuition that the it comes out true – not in the face of such reasonable suspicion about the intuition's credentials.

The example is instructive, for it can easily happen that an account of causation leans heavily on some assumption not explicitly about causation – maybe it's about the truth-value of some conditional, or the identity conditions for some event, or something else – but where, on inspection, there's no obvious way to vindicate that assumption without appealing to the very causal facts the account is meant to treat.

Such assumptions need to be exposed, for they threaten the viability of any account that claims to give *reductive* conditions for causation.

The second sort of bad habit involves the misuse of "pragmatics" in defending an account of causation. One perfectly *justifiable* move is familiar from Mackie (1965), where we gloss certain events or other feature of a case as part of the background context. Lightning strikes a barn, causing a fire. It seems perfectly right, in typical contexts, to call the lightning a cause of the fire, and label the presence of oxygen a mere "background condition". Of course, if we changed the explanatory context, say to one where a middle school science teacher is lecturing her students on how combustion occurs, it could be perfectly right to say that the presence of oxygen *was* a cause of the fire. The change in context moves an event from the background into the foreground, hence changing what counts as an appropriate or correct causal explanation of the case.

But while changing contexts can change what we should mention in an explanation, it does not change the basic causal facts. In our example of the fire, no matter what the context, the domain of causal facts includes *all* the causal facts (e.g., the fact that presence of oxygen is among the causes of the fire), whether or not it is always explanatory or contextually appropriate to refer to some of these facts.

That sort of benign appeal to pragmatics, however, must not be conflated with a much more controversial appeal, one which takes pragmatics to somehow apply to ontology itself. Consider this passage from Pearl and Halpern (2005), where they suggest that facts about what causes what are themselves *model-relative*:

> According to our definition, the truth of every claim must be evaluated relative to a particular model of the world; that is, our definition allows us to claim only that C causes E in a (particular context in a) particular structural model. It is possible to construct two closely related structural models such that C causes E in one and C does not cause E in the other. Among other things, the modeler must decide which variables (events) to reason about and which to leave in the background. We view this as a feature of our model, not a bug. It moves the question of actual causality to the right arena—debating which of two (or more) models of the world is a better representation of those aspects of the world that one wishes to capture and reason about.

The passage suggests two very different ideas – and given the paper's unclarity about just what the "aspects of the world" are that structural models aim to "capture", it's impossible to tell which is in play. One idea – a little surprising, maybe, but all the same quite compatible with the reductionist perspective we're exploring – is

that there is a perfectly objective causal structure in the world that models need to be faithful to, but that a typical model will only *partially* represent this structure; what's more, what is called "causation" in ordinary (and perhaps even scientific) contexts depends on which aspects of the underlying structure are being highlighted by one's choice of model. That's of a piece with the familiar point that we sometimes, for pragmatic reasons, relegate causes to the status of "background conditions". But the second idea is far more radical: it is that *what the world's causal structure is* is somehow relative to one's choice of model – which choice is, evidently, to be made on broadly pragmatic grounds, as witness the reference to "those aspects of the world that *one wishes to capture and reason about*".

Whether or not that view is coherent, it's not compatible with pursuing an ontological reduction. What's more, it is clearly a mistake to think that approaches to causation that relativize causal structure to a choice of model could have widespread application in the natural or social sciences, or indeed, in legal or historical narratives that take themselves to be making factive causal claims. To look at an example, consider recent sociological research suggesting that female applicants whose personal details indicate that they have children are less likely to be offered job interviews, are ranked lower in competence, and are held to a stricter performance standard than male applicants with identical applications (Correll, Benard and Paik 2007). How is it helpful to be told that whether, in fact, being a female-with-children *is* a cause of such discrimination depends on one's choice of model? The obvious retort is that, well, we would like to choose that model which *gets the causal structure right*. But on the second, more radical interpretation of what Pearl and Halpern are up to, there *is no* objective causal structure *to* get right. That won't do. Being able to make objective claims about the causal structure of the world is just too essential to the role actual causal models play in science, and to the ways science is drawn upon, e.g., to develop or encourage governmental and corporate policy.

What we'll take up next are the rules we think should be followed in constructing an ontological reduction of causation to more metaphysically basic facts.

§3.2 The Book of Rules

We describe five of the most important rules.

§3.2.1 Rule one: Thou shalt not smuggle the causal in with the basic.

A traditional conceptual analysis of causation cannot make use of explicitly causal concepts, such as the concept of "intervention" or "manipulation". In a similar fashion, we cannot successfully reduce causal facts to ontologically more basic facts if those facts *include* causal facts. This is obvious. It should be equally obvious that one

cannot use in one's account notions or facts that are merely implicitly or indirectly causal.

Transgressions of this rule can be subtle. Suppose you think that causal facts are to be analyzed, in Davidsonian fashion, as instances of "covering" laws. Does your analysis need, in order to be extensionally adequate, a distinction between *causal* and *non-causal* laws? If so, it violates rule one. Or suppose that your analysis makes use of counterfactuals – but you turn around and give these *causal* truth-conditions. We'll periodically have occasion, in what follows, to highlight points at which an account undercuts its reductive aspirations in just this way.

§3.2.2 Rule two: Thou shalt not be metaphysically extravagant.

You can undercut the explanatory value of your account of causation by characterizing the ontologically basic facts that serve as its ingredients in too metaphysically extravagant a fashion. Example: Suppose your account accommodates causation by omission. But it does so in part by positing a special kind of "negative" event in Meinongian style. Thus, when Billy's failure to water Suzy's plants causes their death, that is in virtue of a relation between one such negative event – Billy's failure to water the plants – and another, more prosaic event (the plants' death). Negative events must not, you insist, be identified with ordinary, "positive" events. You proceed to construct a Grand Metaphysical Theory of them, in order to answer such questions as these: Where do they take place? How long do they last? When are they identical to or distinct from one another? What are their parts?

Regardless of how clever you are in constructing your theory, something has gone wrong. We started with something metaphysically prosaic, and ended up trying to illuminate its nature by appeal to something else that is far too metaphysically extravagant. Now, what *counts* as metaphysical extravagance is to some extent relative to a domain of enquiry and to some extent a matter of taste. Still, we think that in the context of contemporary discussions of causation, the standards are reasonably high. Here is a good rule of thumb: the basic ontology needed for causation should not exceed that needed for the fundamental truths of physics. This rule of thumb strikes us as especially appropriate, given that one of the aims of an ontological reduction of causation is to produce something useful to, and illuminating of, scientific practice. (But don't misunderstand us: there might be quite a lot of basic ontology needed for the fundamental truths of physics.) As we'll see, even with this relatively generous interpretation of the rule, we'll come across several examples of accounts that fail to apply it appropriately.

§3.2.3 Rule three: Thou shalt not rely upon explanatorily idle notions.

A notion is *idle* in our sense if, in order to reduce or to fully explicate *it*, one would have to appeal to machinery that would *already* suffice to analyze causation, without any detour through the notion in question. We'll draw on Davidson again for an example. Suppose you think that what it is for *C* to be a cause of *E* is for there to be some law that "covers" *C* and *E*, under suitably chosen descriptions. *C* is an F-event, *E* is a G-event, and *E* is R-related to *C*; and there is a law that says that every F-event is followed by a G-event R-related to it: *that* is the kind of "coverage" that your account takes to be necessary and sufficient for causation. Fine. But the "law" in question is almost certainly not a law of fundamental physics. (Suzy's throw caused the bottle to break. What was the "covering" law? Presumably, something like this: every throw in such-and-such circumstances is followed by a breaking with such-and-such features. That's not a law of fundamental physics.) So you will, at the end of the day, need some account of what it is for this kind of "higher-level law" to obtain. Imagine that you provide an account in terms of certain kinds of counterfactuals – and that it is clear on inspection that you could have applied those counterfactuals *directly* to the analysis of causation. Then you will have broken rule three.

§3.2.4 Rule four: Thou shalt not be an ontological commitment wimp.

One way to avoid breaking any of the foregoing rules is to say very little. If ontological reduction is genuinely one's aim (as opposed, say, to mere necessary connection; see §3.1.1), then all one gains by saying very little are gains of theft over honest toil.

There is a straightforward way *not* to be an ontological wimp: show, explicitly, how facts about causation are grounded in facts about fundamental physical states, together with facts about the fundamental physical laws governing their evolution. Granted, that's a tall order. But keeping it firmly in mind, if only as an ideal, helps guard against the overly cavalier use of unexplained concepts that we will occasionally witness in the accounts we will discuss.

There is a second way to be a wimp about ontological commitments: appeal, in one's reduction, to facts too specific to our own world. Now, we need to be a little careful here. It's not that we think an account needs to be in the business of issuing verdicts about any old conceivable situation some philosopher can cook up. The literature (including some of our own contributions!) occasionally likes to speculate about the causal structure of worlds in which magical spells can act across a temporal gap, or in which there is backwards causation, etc.; part of the lesson of the little case study concerning the direction of time rehearsed in §3.1.4 was to remind us that it's far from obligatory for an ontological reduction to extend its scope so far. So a high-quality account that cannot, alas, say anything coherent about backwards causation and other esoterica should not, for that reason, lose our respect. Still, causation is a

generic enough relation (and our corresponding concept of it is broad enough) that tying a theory of it too closely to facts peculiar to our actual world, and its physics, manifests a failure of nerve. Accounts that do that *should* lose our respect.

§3.2.5 Rule five: Thou shalt not take thine own intuitions too seriously.

Suppose you are interested in what grounds the direction of time. Someone comes along, insisting that it cannot be grounded in the direction of global entropy increase, because such a view fails to do justice to their intuitive conviction that time genuinely *passes*. Or suppose you're curious about what it is for an object to be *solid*. Someone comes along, insisting that there are in fact no solid objects, since atomic theory shows that most of the things we mistakenly take to be solid are composed largely of empty space – and it is *intuitively clear* that if an object is solid and occupies a certain region R of space, than for any subregion of R, some part of that object occupies that subregion. What should you think?

In both cases, you should think that intuition has been set up as an arbiter of questions it is not competent to judge. Granted: when investigating some aspect of the ontological structure of the world, it is hugely important to pay attention to ordinary intuitions as a valuable source of clues for where to look. But the process of theorizing can yield ample opportunities for rejecting some of these intuitions as misguided (though it will help, if we can supplement the theory in question with an explanation for why we were so easily led astray). This methodological point is blindingly obvious in the case of the direction of time, or the nature of solidity. Thankfully, the literature on causation is beginning to incorporate it as well. It is no longer so acceptable to claim, with Lewis (1986b), that "If an analysis of causation does not deliver the common-sense answer, that is bad trouble." In sum: while you should certainly worry that your analysis has missed something important if it flouts some clear and firmly held intuition, you should not hastily assume that your analysis has been refuted.

§3.3 The accounts reconsidered

We're now in a good position to reconsider the accounts sketched in §2. We'll highlight the most important issues.

§3.3.1 Regularity accounts and laws of nature

Recall that we distinguished two varieties of regularity account. One takes as its key idea that for *C* to cause *E* is for these events to be *covered* (perhaps under appropriate descriptions) by a suitable *law*. The other says that *C* must belong to a set of conditions minimally sufficient for *E*. The second idea brings laws in only indirectly,

to say what "sufficient" means. And, as we saw, the laws brought in can simply be the laws of fundamental physics.

What about the first idea? A dilemma confronts it. Maybe the covering laws are supposed to be *fundamental* laws. But these laws relate, in the first instance, *complete physical states* of the world to subsequent complete physical states. We will search in vain among them for laws that will explicitly "cover" any but an exceedingly narrow range of causal phenomena. Suzy throws a rock at a bottle, shattering it; are we to suppose that there is some way of describing her throw and the shattering such that, relative to these descriptions, the relation between the two events can be seen as an instance of some *fundamental* law? On the other hand, maybe the laws are the far-from-fundamental laws of the special sciences. But then the account will almost certainly flout one or more of our rules. It might flout rule one by taking the "laws" simply to *be* certain kinds of causal generalizations. Or rule two, by treating them as sui generis, and irreducible to more basic physical laws. (Cartwright 1999 seems to have a view that is something like this.) Or rule three, by analyzing these laws in counterfactual terms themselves adequate to analyze causation. Or rule four, by saying nothing about what these laws come to (Maudlin 2004, though in other respects quite brilliant, is an example).

We think the best way to avoid the dilemma is to abjure the style of regularity account that gives rise to it; that is why we favor, as a more interesting and fruitful approach, the second of the two ideas sketched in §2.1.

§3.3.2 Truth conditions of counterfactuals

A successful account of causation cashed out in terms of counterfactuals needs a successful account of counterfactuals, one not itself cashed out in causal terms. (Note that this point applies not merely to the accounts sketched in §2.2, but also to the "updated regularity account" sketched in §2.1.) So not all counterfactual treatments of causation will be suitable candidates for an ontological reduction. (For example, Woodward's 2005 nonreductive approach will not meet these standards.) One might think that there is not far to look – Lewis's oft-cited 1979 analysis aims to provide just what we need. But there are serious problems with Lewis's view.

Recall what Lewis's analysis says about a simple counterfactual of the form "if event *C* had not occurred, then event *E* would not have occurred". Assuming that *C* and *E* in fact occur, we evaluate this counterfactual by moving to a possible world with the following features: Up until shortly before the time of *C*'s occurrence, its history is *perfectly qualitatively identical* to the actual history. And then a "miracle" occurs – a violation of the actual fundamental laws of nature (though, obviously, not a violation of the laws that hold in the counterfactual world itself). Post-miracle history unfolds perfectly in accordance with the actual laws. The miracle should be as small

and as inconspicuous as possible, subject to the requirement that it throws history off course enough to make it the case that *C* does not occur. The counterfactual is true, finally, just in case *E* also fails to occur.

We think that this is close to the account of the truth conditions for counterfactuals that a philosopher should endorse, who wants to give an ontological reduction of causation in terms of counterfactuals. But it's not quite right, and, more importantly, the motivations behind it strike us as deeply flawed. To clarify all this, let's look a bit more deeply into what Lewis took himself to be trying to accomplish in providing truth conditions for counterfactuals, and why he thought these truth conditions would take the form of the foregoing "small-miracles recipe" for the specific sorts of counterfactuals that appear in his analysis of causation.

Obviously, one of the aims Lewis had was to meet the needs of his counterfactual analysis of causation, and to do so in a suitably reductive manner (so that the proffered truth conditions for counterfactuals did not themselves make use of any causal notions). But he also took on board, more or less explicitly, three additional constraints:

First, he took it for granted that his truth conditions should be *general purpose*, and not simply tailored to the kinds of counterfactuals needed in his analysis of causation. Thus, these truth conditions should be able to handle sentences such as "if kangaroos had no tails, then they would fall over", and even "if gravity worked by an inverse-cube law, planetary orbits would still sweep out equal areas in equal times".

Second, he took for granted that the proposed truth conditions should fit within a general framework of *similarity semantics* for counterfactuals. That is, we start with the assumption that the right form for the truth conditions for sentences of the form "if A were the case, then B would be the case" ("A→ B", for short) is roughly as follows: among those possible worlds in which A is true, the one that is most similar to the actual world is one in which B is true.[17] The project then becomes to articulate the specific standards of similarity that our counterfactuals implicitly make use of.

Third – and rather too ambitiously – Lewis wanted an account of counterfactuals that would explain the asymmetry of time, in line with the idea that what distinguishes the future *as such* is that it counterfactually depends on the past, but not vice versa.

Lewis managed to leverage these constraints into a specific standard for similarity, one that, no matter how much one admires the cleverness of its construction,

[17]We say "roughly" because there are complications if – as will surely typically be the case – there is no uniquely most similar A-world.

ought also to strike one as hopelessly byzantine. Here it is: In selecting a most similar A-world, it is of first importance to avoid large, widespread miracles, of secondary importance to maximize the region of exact match of particular facts, of third importance to avoid small miracles, and of little importance to secure approximate match. So the general-purpose truth conditions are simply these: A→ B is true just in case, among those possible worlds in which A is true, the one that is most similar to the actual world *according to the foregoing standard* is one in which B is true. Lewis argues that, from these generic truth-conditions, one can derive the small-miracles recipe for evaluating causal counterfactuals. Along the way, we get an argument for why, in worlds like ours, the future typically counterfactually depends on the past, but not vice versa.

But however influential, the account is wholly unsatisfactory, for reasons that by now are well known. Here are three serious problems. First, Adam Elga has argued decisively that given what we know about thermodynamics and statistical mechanics, Lewis's claim to have shown where the asymmetry between past and future comes from cannot be sustained. Second, the standard of similarity between worlds that is supposed to yield the small-miracles recipe seems far too convoluted not to be viewed as ad hoc and arbitrary. Among all the standards of similarity that we might have hit upon to govern our use of counterfactual conditionals, what could explain our choice of *this* one? (See Horwich 1993 for a forceful presentation of this objection.) Third, the small-miracles recipe almost always produces a gap between the time of occurrence of the miracle in the counterfactual world, and the time of the event in question – in which case there will be counterfactual dependence of the immediate past *on the present*. Woodward (2005) argues persuasively that this there is no way to avoid the result that – if counterfactual dependence suffices for causation – backwards causation is rampant.

And anyway, there is another issue to consider: Why should those of us whose reductive ambitions are focused squarely on causation – and causation alone – endorse all the elements of the Lewisian semantics? If our first interest is in the prospects for a counterfactual analysis of causation, then the right thing to do is not to try to fix Lewis's account, but rather to abandon it – and more to the point, to abandon the pretensions to providing truth conditions for counterfactuals that will simultaneously serve the needs of a theory of causation *and* meet Lewis's additional constraints. It should not be thought that abandoning these pretensions comes at any cost, or is any occasion for disappointment. Remember that what we are aiming at is ontological reduction. We are not trying to uncover the structure of our ordinary concept of causation (except insofar as doing so provides us clues as to where interesting ontological reductions might be found), nor are we trying to uncover the con-

nections between this concept and our ordinary concept of counterfactual dependence. So we can, with a perfectly clear philosophical conscience, look for an account of the truth conditions of counterfactuals that tries to do nothing more than give us a useful tool in the construction of an account of causation. And once we limit our focus in this way, the needed account is not hard to find.

Here is the simple alternative to Lewis's account (adapted from Maudlin 2007b) that we have in mind. Suppose event *C* occurs at t, and event *E* occurs later. To evaluate "if *C* had not occurred, then *E* would not have occurred", we construct a counterfactual state of the world at time t as much like the actual state at time t as possible, save for the fact that *C* does not occur. Think of taking the actual time-t state of the world, and ringing carefully localized changes on it just sufficient to make it the case that *C* does not occur. (An important refinement of this procedure will appear shortly.) We then evolve the resulting state forward, in accordance with the actual laws of nature. If the resulting history yields *E*, the conditional is false; otherwise it is true. That's the recipe – tailored, as you can see, to the kinds of counterfactuals needed in a theory of causation.[18] We *don't* try to display this recipe as an instance of some more general prescription for evaluating counterfactuals. Similarity enters in not as a relation between whole *worlds*, but as a relation between *states* of worlds at times. We *don't* try to get to the counterfactual state in which *C* fails to occur by way of some miracle that throws history off course – in fact, we don't care one whit where this state came from. (Thinking in terms of such history-altering miracles might be a psychologically useful method for fixing attention on the appro-

[18] Note that there is an issue here derived from statistical mechanics that we are overlooking. Consider a simple example. A gas is confined by a barrier to one half of a chamber, the other half of which contains a vacuum. Suppose we ask what would have happened if, at time t, the barrier had been removed. We would like to say that the gas would have diffused across to the other side of the chamber. And our recipe seems to guarantee this result: we construct a counterfactual state of the world at time t which is just like the actual state, save that the barrier is absent; evolving this state forward in time would seem to yield a future in which the gas diffuses. But that is not quite right. The more accurate picture is really this: There is not one single counterfactual t-state that meets our conditions, but rather a continuous infinity of such states, differing in minute and seemingly insignificant microphysical respects (in one, a certain gas molecule is moving with just this velocity; in another, it is moving with a slightly different velocity; etc.). But some of those states will be bizarre anti-entropic states, that yield forward evolutions in which the gas stays on one side of the partition. We know, on statistical mechanical grounds, that these bizarre states make up an astronomically tiny minority of all of the relevant counterfactual states. But we also know that they exist. So what we should really say is this: if the partition had been absent at time t, then, with a statistical probability vanishingly close to but not exactly equal to 1, the gas would have diffused across to the other side of the chamber. We will take it for granted henceforth that counterfactuals like this are good enough for the purposes of an ontological reduction of causation.

priate counterfactual state; to this extent, the attractiveness of the small-miracles recipe makes sense.) And, finally, we don't try to squeeze out a story about the direction of time from our analysis of counterfactuals.

This alternative to the Lewis analysis is the one that we will make use of henceforth. Let's see it in action, applied to the example depicted in figure 1:

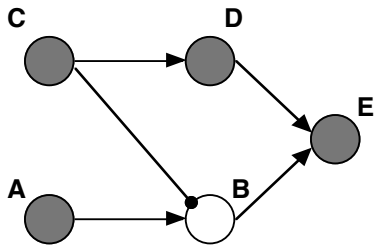

Figure 1

At time 0, neurons **C** and **A** both fire. To show that *E* does not depend on *C*, we simply change this time 0 state in a localized way, making neuron **C** dormant. The events that unfold are those depicted in figure 2:

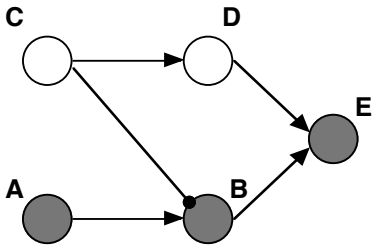

Figure 2

At time 1, neurons **D** and **B** both fire. To show that *E* in figure 1 *does* depend on *D*, we simply construct a counterfactual state for time 1 in which neuron **D** is dormant, and everything else remains the same – so **B** is *not* firing, and the neurons are still connected together in the way depicted in figures 1 and 2. We simply don't *care* where this state came from – if you like, imagine that the counterfactual world just *starts out* in this state. It is then clear that the counterfactual state unfolds in such a way that **E** does not fire. What this result illustrates is that our simple alternative recipe for evaluating counterfactuals has the "non-backtracking" feature that, as we saw

back in §2.2.1, it needs to have in order to have a chance of undergirding a successful account of causation.

§3.3.3 Default and deviant states

We have one more serious issue to deal with, one that will likely arise for *any* account of counterfactuals. (It certainly arises both for Lewis's miracles-based recipe, and for our own "altered states" recipe.) Even the simplest of examples illustrates it. Suzy throws a rock at noon, breaking a bottle. It is utterly natural – and surely correct – to hold that if she hadn't thrown the rock, the bottle would not have broken (for remember that this is not one of those tricky cases in which some backup process aims to break the bottle as well). But then we must be supposing that, in the relevant counterfactual situation in which Suzy is not, at noon, throwing a rock at the bottle, she is not doing anything *else* that would lead to a bottle-breaking: she is not starting to run up towards the bottle to level a kick at it; she's not throwing some other object at it; she's not shooting her slingshot at it; etc.

Neither the small-miracles recipe that Lewis favors, nor the altered-states recipe we favor, automatically secures this result. Our own recipe instructs us to construct a counterfactual state of the world at noon by taking the actual state, and locally modifying it so that Suzy does not throw a rock. But – given the vast multitude of ways she could turn out to *not* be throwing a rock – these instructions underspecify what she is doing *instead*. It would be foolish to appeal to similarity here, as if the right way to proceed is to have her do something *very much like* throwing a rock. The small-miracles recipe is, if anything, in even worse shape. Suppose that, shortly before noon, Suzy is deliberating about the best way to break the bottle. In fact, she settles on throwing a rock, rather than firing her slingshot (her second choice). If that is how things play out, then the smallest, most inconspicuous miracle that will throw history off course just enough to get her not to throw her rock will consist in a few subtle alterations of the neural underpinnings of her deliberations, alterations that lead her to fire her slingshot instead. So whereas our altered-states recipe wasn't fleshed out enough to yield a determinate result, Lewis's small-miracles recipe is sometimes guaranteed to yield the *wrong* result.

This problem has been noticed before – for example, by Lewis himself. Here's a pithy expression of it:

> What is the closest way to actuality for *C* not to occur? – It is for *C* to be replaced by a very similar event, one which is almost but not quite *C*, one that is just barely over the border that divides versions of *C* itself from its nearest alternatives. But if *C* is taken to be fairly fragile [i.e., characterized by stringent conditions of occurrence], then if *C* had not occurred and almost-*C* had occurred instead, very likely the effects of almost-*C* would have been much the same as the actual effects of *C*.

> So our causal counterfactual will not mean what we thought it meant, and it may well not have the truth value we thought it had. When asked to suppose counterfactually that *C* does not occur, we don't really look for the very closest possible world where *C*'s conditions of occurrence are not quite satisfied. Rather, we imagine that *C* is *completely and cleanly excised from history*, leaving behind no fragment or approximation of itself. (Lewis 2004a, p. 90; italics added)

We think Lewis's observations are right on target – up to the italicized portion, at which point they become mysterious. What exactly does such "complete and clean excision" consist in? Removal of the event by some sort of metaphysical scalpel? Leaving behind … what? The Void? (We should also note that it is unclear how Lewis's approach fits with his theory of event essences and causation in his 1986c.)

A much better view is that for any given event, we work with an antecedently understood distinction between a *default state* for the region in which the event occurs, or for the physical system or systems to which it pertains. Conceiving of the event as one among various possible *deviations* from that default state, we answer the question, "What would have happened, had that event not occurred?" by returning the relevant region or system to its default state, holding the state of everything else fixed. It is in this way – and not by metaphysical surgery – that we can fill in the altered-states recipe for evaluating counterfactuals. Thus, the counterfactual noon-state we have in mind in the world where Suzy does not throw her rock is one in which she is standing idle, doing nothing.

We strongly suspect that any ontological reduction of causation that makes use of counterfactuals will need to deploy some distinction between default states and deviations thereof. That seems obvious in the case of Lewis's original analysis, though perhaps less so in the case of his influence analysis, or de facto dependence approaches. The need is even more glaring in the case of the second of our two regularity accounts – which, remember, analyzed what it is for a set of events S, all occurring at time t, to *suffice* for some later event *E* by means of the conditional, "if only the events in S had occurred at t, then *E* would still have occurred". Clearly, understanding this conditional requires an understanding of what it comes to for nothing *else* to be happening at the relevant time – which looks to be the same as saying: everything else is in its default state.

At any rate, we won't argue the point further. We simply note that if our suspicion is right, then a major piece of unfinished business for ontological reductions of causation that make use of counterfactuals is to provide a supplementary account of this distinction. Moreover, such an account needs to respect the reductionist constraints we've laid out above. This is an area in which, at present, matters are very

much wide open (see Maudlin 2004, Hitchcock 2007, and Hall 2007 for some discussion and attempts to apply a default/deviant distinction).

One final note. It seems to us that the widespread use of neuron diagrams is partly to blame for the fact that philosophical discussions of causation typically overlook the centrality of the default/deviant distinction. And that is because, in the case of neurons, it is so obvious as to escape notice what counts as the relevant default state for a neuron: it is just the *dormant* state.

§3.3.4 De facto dependence counterfactuals

So far we have seen that – modulo some lingering and perfectly legitimate concerns about the status of the default/deviant distinction – we have been able to give suitably non-causal truth conditions for one important kind of counterfactual. That covers a lot of territory. But not all of it. What about the sorts of counterfactuals needed in de facto dependence accounts? Recall the general form of these accounts: what it is for event *C* to be a cause of another event *E* is for it to be the case that, for some suitably chosen fact F about the given situation, if *C* had not occurred but F had still obtained, then *E* would not have occurred. It is challenging to specify how the fact to be held fixed gets selected. Let us set that issue aside for now, and focus on the question of how, once the fact F has been selected, truth conditions are to be given for this counterfactual.

Everything depends on the form that F takes. If it takes the right form, then truth conditions come easily, by way of a natural extension of our altered-states recipe. But if it takes the wrong form, then it remains entirely too obscure what these truth conditions are. We will illustrate these two possibilities by a pair of examples.

Start with the case where things work well. Consider figure 1 again. Suppose, as part of your de facto dependence account, you have identified the fact to be held fixed as the fact that neuron **B** does not fire at time 1. Then it is easy to extend our altered-states recipe in a way that allows for the clean evaluation of the conditional "if **C** had not fired at time 0, but **B** had still failed to fire at time 1, then **E** would not have fired". We do so as follows:

First, focus on the actual time-0 state of the world. Locally modify it so as to make it the case that **C** does not fire (i.e., return **C** to its dormant state). Evolve the resulting state forward until time 1. The result is a state in which **B** is firing. Now make local changes to *this* state, so as to make it the case that **B** is dormant at time 1 (i.e., in the same state as it is *actually* in at that time). Evolve this resulting state forward until time 2. **E** does not fire. So the conditional is true.

More generally, if we have a conditional of the form "if *C* had not occurred, but the fact F had still obtained, then *E* would not have occurred", and there is a non-arbitrary way to make the fact F obtain by locally modifying the state of the world at

one or more times, then we can follow the same procedure: modify the state of the world at the time at which *C* in fact occurs so as to make it the case that *C* does not occur; update in accordance with the actual laws; and make localized modifications along the way, in a non-arbitrary fashion, so as to guarantee that fact F still obtains. If the fact F simply consists in the occurrence or nonoccurrence of specific, localized events, then this will in general be straightforward.

So far, so good. Unfortunately, not every case will be like this. Recall the example from §3.1.5, which resists such a clean treatment: Suzy and Billy both throw rocks at a bottle, but Suzy's gets there first, shattering it. If she had not thrown, then Billy's rock would have shattered the bottle a moment later. It is a commonplace among fans of the de facto dependence approach to causation to point out that Billy's rock never in fact strikes the bottle, and to go on to claim that the conditional that grounds the fact that it is Suzy's throw that causes the bottle to break is therefore this one: if Suzy had not thrown, and Billy's rock had still somehow failed to strike the bottle, then the bottle would not have broken. But we immediately run into trouble if we try to analyze this counterfactual in the way just indicated. The fact to be held fixed is too indeterminate for us to be able to tell just which state of the world to locally modify, so as to guarantee that this fact still obtains in the relevant counterfactual situation. Worse: some local modifications will get exactly the wrong result – for example, the local modification that puts the bottle into a shattered state before Billy's rock can reach it.

As yet, there is no appropriately reductive account of the truth-conditions of de facto dependence counterfactuals that we know of that deals with this problem.

§3.3.5 Conserved quantities

While we treat transference accounts as live options, we want to register two very serious complaints. The problems we have in mind are not tied to any particular example, but have much more to do with a failure to abide by the methodological precepts we think should guide philosophical inquiry into causation.

First, transference accounts seem to suffer from a surprising lack of ambition. (Cf. our rule, "thou shalt not be an ontological commitment wimp".) Even if these views correctly describe the actual world, surely there could be worlds with laws that don't single out anything as a "conserved" quantity – more generally, that do not describe the transfer of *anything* physically fundamental. Consider, for example, a world described in Maudlin (2004) that operates on principles akin to those at work in Conway's game of "Life": space is divided up into discrete cells, each of which can be either occupied or unoccupied; time is divided up into discrete moments; the pattern of occupation of the cells at one moment is lawfully and deterministically fixed by the pattern of occupation of the cells at the prior moment. There seem to be

causal relations in such a world, and we are perfectly capable of recognizing them. It is a mark against transference accounts that they can have nothing to say about why this is so.

But there is a more serious problem. Let us illustrate it by means of our example of Billy, Suzy, and the bottle. Suzy's throw is a cause of the bottle's shattering and Billy's is not. Can a transference account illuminate why this is so? You might think so. After all, it is Suzy's rock, and not Billy's, that transfers momentum or energy to the bottle, isn't it? To wit, consider what Ehring says about such cases: "Causal ancestry is determined by the origins of the energy/momentum manifested in the effect. A preempting cause is distinguishable from a preempted cause in virtue of the fact that the energy/momentum of the effect-event is traceable back to the preempting cause-event, but not to the preempted cause-event" (Ehring 1997, p. 45).

Such an analysis relies on too soft a focus. Consider that Billy's rock, as it flies through the air, pushes air molecules ahead of it, and that some of these bump into the bottle before Suzy's rock strikes it. We can, for that reason, credit Billy's throw with initiating a process that transfers energy, momentum, or indeed any other candidate quantity to the bottle. That is, whatever the stuff is whose transfer to the bottle makes it the case, according to a transference account, that Suzy's throw causes the bottle to break, it seems that we can find that quantity transferred to the bottle by Billy's throw as well. This creates trouble for the transference theorist, whose view seems to entail that "[i]f there is a transfer from both the main and the alternate lines, then there is simply no preemption, but only two lines of partial contributing causes" (Ehring 1997, 45).

Now, what one obviously wants to say is that whereas Billy's throw might transfer momentum (for example) to the bottle, it does not transfer *enough* to make the bottle shatter. So the limited amount of momentum Billy's throw transfers is not sufficient to render it causally relevant to the breaking. That is perfectly correct, but what transference accounts fail to do, as far as we can tell, is to provide any illumination about *why* it is correct. What's more, it is fairly obvious where such illumination should come from: We might, for example, focus on the fact that the breaking does not *counterfactually depend* on the transfer of such a small quantity of momentum (whereas, by contrast, it *does* depend on the transfer of the larger quantity of momentum that resulted from Suzy's throw); or we might focus on the fact that the transfer of the smaller quantity is not *sufficient* in the circumstances for the shattering (whereas the transfer of a larger quantity is). That is, we would focus on the kinds of relations that counterfactual and regularity accounts place at center stage.

This sort of problem is going to be ubiquitous (unless, perhaps, we choose to restrict our attention to causation among the most microphysical events we can find).

It seems to us very likely that transference accounts can have a chance of solving it only if they incorporate analytical tools – maybe counterfactual dependence, maybe some notion of sufficiency – that can be independently used to provide an account of causation. If so, pure transference accounts inevitably violate rule three ("thou shalt not rely upon explanatorily idle notions").

§4 Concluding remarks

We hope you share our enthusiasm concerning the value and interest of ontological reductions in general, and of causation in particular. But maybe you don't. That could be because you're just interested in other parts of philosophy. No worries. But it could also be because you are laboring under one or another misconception. Two such misconception are especially worth exposing.

The first goes like this: "What the point in continuing to pursue an analysis of causation? We've been at it, like, *forever* – and all that's happened is that ever more baroque analyses confront ever more baroque counterexamples. We should give up, and do something productive with our time."

However common this attitude (in some circles, anyway), it doesn't sustain critical scrutiny. To begin, the best of the going analyses are really not *that* baroque. But there's a deeper confusion, which is that the name of the game ought to be to construct an analysis *that successfully runs the gamut of all possible counterexamples*. We agree: *that* game isn't particularly worth playing. (What exactly would you have gained, if you succeeded at it?) But we hope that our discussion in this essay has made it patently obvious that someone pursuing an ontological reduction has very different aims. You can't completely *ignore* intuitions about cases, in pursuing these aims. But you're not trying to triangulate to them, either.

The second misconception is that a successful reductive account of causation wouldn't yield anything of value. Now, we think this complaint can in fact be answered on its own terms: reductive accounts hold great promise in clarifying, for example, the relationship between statistical correlation and causation; they also have helped clarify the status of "laws" in the special sciences. But it's really better just to reject the terms themselves. Our world has, somehow, a rich causal structure. A philosopher pursuing an ontological reduction of causation wants to understand what, fundamentally, this structure consists in. It would seem ample motivation for such a project that one is, simply, *curious*.

§5 References

Cartwright, Nancy 1999: *The Dappled World*, Oxford: Oxford University Press.

Chalmers, David 2002. "Does Conceivability Entail Possibility?", in T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility*: 145-200.

Chalmers, David 2011. "Verbal Disputes", *Philosophical Review*, 120:4.

Collins, John; Hall, Ned; and Paul, L. A. eds. 2004a. *Causation and Counterfactuals*, Cambridge, MA: MIT Press.

Correll, S. J., Bernard, S., & Paik, I. (2007). "Getting a job: Is there a motherhood penalty?", *American Journal of Sociology* 112: 1297-1338.

Davidson, Donald 1967. "Causal Relations", *Journal of Philosophy* 64: 691-703.

Dowe, Phil 2000. *Physical Causation*. New York: Cambridge University Press.

Eells, Ellery 1991. *Probabilistic Causality*. Cambridge: Cambridge University Press.

Ehring, Douglas 1997. *Causation and Persistence*, New York: Oxford University Press.

Elga, Adam 2000. "Statistical Mechanics and the Asymmetry of Counterfactual Dependence." *Philosophy of Science* (suppl. vol. 68, PSA 2000): 313-324.

Fair, David 1979. "Causation and the Flow of Energy", *Erkenntnis* 14: 219-50.

Fine, Kit. (2003). "The Non-identity of a Thing and its Matter" *Mind* 112 (446), 195-234.

Frick, Johann 2009. "'Causal Dependence' and Chance: The New Problem of False Negatives", ms.

Gopnik, A. Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T, & Danks, D. (2004). A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review* 111: 1-30.

Hall, Ned 2007. "Structural Equations and Causation", *Philosophical Studies* 132: 109-136.

Hall, Ned and Paul, L. A. 2011. *Causation and the Counterexamples: A User's Guide*, forthcoming from OUP.

Halpern, Joseph and J. Pearl 2005. "Causes and explanations: A structural-model approach-Part I: Causes", *British Journal for the Philosophy of Science* 56:843-887.

Hitchcock, C. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs", *Journal of Philosophy* 98: 273-299.

Hitchcock, Christopher 2007. "Prevention, Preemption, and the Principle of Sufficient Reason", *Philosophical Review* 116: 495-532.

Horwich, Paul 1993. "Lewis's Programme", in E. Sosa and M. Tooley (eds) 1993, *Causation*: Oxford University Press.

Kvart, Igal 2004. "Causation: Probabilistic and Counterfactual Analyses", in Collins, Hall, and Paul (eds), 2004.

Lewis, David 1973a. "Causation", *Journal of Philosophy* 70: 556-67. Reprinted in Lewis 1986a: 159-172.

Lewis, David 1979. "Counterfactual Dependence and Time's Arrow", *Noûs* 13: 455-476. Reprinted with Postscripts in Lewis 1986a: 32-66. Citations are from the latter printing.

Lewis, David 1986a. *Philosophical Papers, Volume II*, Oxford: Oxford University Press.

Lewis, David 1986b. "Postscripts to 'Causation'", in Lewis 1986a: 172-213.

Lewis, David 1991. *Parts of Classes*. Blackwell.

Lewis, David, 2000. "Causation as Influence", *Journal of Philosophy* 97: 182-197.

Lewis, David, 2004a. "Causation as Influence", in Collins et. al (eds) 2004; this is an expanded version of Lewis 2000.

Loewer, Barry 1996. "Humean Supervenience", *Philosophical Topics* 24: 101-127.

Lombrozo, Tania 2010. "Causal-explanatory pluralism: How intentions, functions, and mechanisms influence causal ascriptions." *Cognitive Psychology* 61, 303-32.

Mackie, J.L. 1965. "Causes and Conditions", *American Philosophical Quarterly* 2: 245-264.

Maudlin, Tim 2004. "Causation, Counterfactuals, and the Third Factor", in Collins, Hall, Paul (eds) 2004; reprinted in Maudlin 2007a.

Maudlin, Tim 2007a. *The Metaphysics Within Physics*, Oxford: Oxford Univ. Press.

Maudlin, Tim 2007b. "A Modest Proposal Concerning Laws, Counterfactuals, and Explanation", in Maudlin 2007a.

Maudlin, Tim 2007c. "On the Passing of Time", in Maudlin 2007a.

Nickel, Bernhard 2008. "Generics and the Ways of Normality", *Linguistics and Philosophy*, 31(6): 629–648.

Pearl, Judea 2009. *Causality: Models, Reasoning and Inference*. Cambridge: Cambridge University Press.

Ramachandran, Murali 2004. "A Counterfactual Analysis of Indeterministic Causation", in Collins, Hall, and Paul (eds) 2004.

Salmon, Wesley 1994. "Causality Without Counterfactuals", *Philosophy of Science* 61: 297-312.

Simons, Peter 1987. *Parts: A Study in Ontology*. OUP.

Sloman, S.A. (2005). *Causal models: how people think about the world and its alternatives*. New York, NY: Oxford University press.

Strevens, Michael 2009. *Depth: An Account of Scientific Explanation*, Cambridge: Harvard Univ. Press.

Wolff, P. (2007). Representing causation. *Journal of Experimental Psychology: General*, 136: 82-111.

Woodward, J. 2005. *Making Things Happen: A Theory of Causal Explanation*, Oxford: Oxford Univ. Press.

Yablo, Stephen 2004. "Advertisement for a Sketch of an Outline of a Proto-Theory of Causation," in Collins et. al (eds) 2004.