The Intrinsic Character of Causation

Ned Hall

§1 Introduction

There is surely something to the idea that the causal structure of a process is *intrinsic* to it—determined, that is, by the intrinsic natures of the events that make up the process, together with the ways in which they are juxtaposed with one another, together with the laws that govern that process. Consider a simple case.

Suzy Alone: Suzy, an expert rock-thrower with a taste for minor acts of destruction, throws a rock at a bottle. The rock hits the bottle, shattering it. Her throw is a cause of the shattering. Isn't it also clear that in addition to the laws, whatever *makes* it a cause of the shattering is to be found wholly within the sequence consisting of the throw, the shattering, and the events in between?[1] Whatever is happening off-stage, as it were, would seem to have no bearing on the causal status of her throw.

As further evidence that some such thesis of Intrinsicness (as I will call it) guides our thinking about causation, consider a variant.

Suzy First: This time, Suzy's friend Billy throws a rock at the bottle, too. He's just as expert as she is, but a bit slower. Consequently, her rock gets there first; but if she hadn't thrown it, the bottle would have shattered all the same, thanks to his throw.

Billy's throw in Suzy First disrupts the neat nomological relationships that held between Suzy's throw and the shattering, in Suzy Alone: For example, the shattering no longer counterfactually depends on her throw, nor does her throw belong to a unique set of non-redundant sufficient conditions for the shattering.[2] No wonder that cases like Suzy First make

[1] Maybe we didn't need to say, "in addition to the laws"—maybe, that is, the fact that the process is governed by such-and-such laws is itself intrinsic to that process. We will come back to this issue later.

[2] Some would object, either on the grounds that the shattering *does* still depend on Suzy's throw—for without her throw the window would have shattered moments later, and that would have made for a numerically different

such trouble for philosophical analyses of causation. But they don't trouble our intuitions in the slightest; far from being a borderline case of causation, Suzy's throw in Suzy First is just as clearly a cause of the shattering as in Suzy Alone (and, equally clearly, Billy's is *not*). Plausibly, what is guiding our judgments about the more complex case is that the extra complexity Billy's throw brings with it is *extrinsic to*—and therefore irrelevant to the evaluation of—the sequence connecting Suzy's throw to the shattering. Think of the matter this way: We begin with the pared-down situation, where Suzy alone is present, and throws. We now *add* elements to the situation—we add Billy, his throw, the flight of his rock toward the bottle—but in such a way as to leave unchanged the intrinsic characteristics of the process that begins with Suzy's throw and ends in the shattering.[3] And so, of course, Suzy's throw remains a cause of the shattering.

The Intrinsicness thesis that I will articulate and explore in this paper aims at codifying the powerful intuitive conviction that such extrinsic additions can make no causal difference. This thesis is valuable for a number of reasons, but the one that I will place front and center is this: it allows us to solve what would otherwise be an intractable problem posed by depressingly simple examples of causal overdetermination. (Suzy First is a typical instance.) Other benefits, which I will only have space to sketch, are these: it provides the ingredients for an elegant account of *symmetric* causal overdetermination, and it helps illuminate the notion of *causal efficacy* for event properties. It also has some striking consequences, for if it is correct then the core idea behind counterfactual analyses of causation is false, and, furthermore, there is no such thing as causation by omission. Of course, those consequences will look rather *too* striking to some: they will seem more accurately labeled "costs". Later, we will see how to restate them in ways that

event—or on the grounds that the nomological relations I have attended to are too simple, and that more intricate ones can be found to distinguish Suzy's from Billy's throw. The first objection is pretty clearly misguided; the second, while obviously too open-ended for decisive refutation, becomes implausible on close inspection. I'll explain all this in §5, below.

[3] Well, *almost* unchanged, and at any rate unchanged in any *relevant* respect. We will examine the need for the qualifications—and how to state them in a way which is not objectionably weaselly—in §8, below.

will make them appear less drastic: namely, by taking Intrinsicness to characterize one central *kind* of causation.[4]

The plan is as follows: After some preliminaries, I will give the formulation of the Intrinsicness thesis I prefer, and then raise several questions about it. Answering these questions will take us through a number of topics. The first concerns the motivation for the Intrinsicness thesis; here I will lay out the philosophical problem posed by causal overdetermination. The second focuses on the reasons why the Intrinsicness thesis needs to have the somewhat complicated form that I will give it. The third concerns the applications of the Intrinsicness thesis just alluded to. It is here that I will argue against counterfactual analyses, and the claim that there is such a thing as causation by omission; I will also offer my solution to the problem of overdetermination, and indicate in passing how to leverage the thesis into an adequate account of symmetrical overdetermination. The fourth topic concerns the need for a certain key refinement of the Intrinsicness thesis; I will sketch what I think is the best way to undertake that refinement—and again, indicate in passing how this refinement helps explain what causal efficacy consists in. Finally, I will close with a brief discussion of an important objection, the reply to which will help to draw out what I take to be a deeper motivation behind the Intrinsicness thesis. To foreshadow: I think that if we wish to have the means for elucidating, in a manageable form, the nomological structure of our world, then we need a causal concept governed by Intrinsicness.

Now to the preliminaries, the first of which aims to forestall confusion about the nature of the Intrinsicness thesis.

§2 Two roads not taken

Some authors would find the statement of the Intrinsicness thesis to be simplicity itself, holding that whenever an event *c* causes an event *e*, this causal relation simply *is* intrinsic to the

[4] This view is developed more fully in my 2003.

pair (*c*,*e*). To be sure, the relation is *external*, in that (like spatiotemporal relations) whether it obtains is not solely determined by the respective intrinsic properties of *c* and *e*. But it is nevertheless intrinsic to the pair—and that is the end of the story.[5]

This position is most naturally developed as part of a certain kind of *non-reductionist* position about causation, according to which facts about what causes what are metaphysically primitive (Tooley 1990). To see why, start with what must seem an obvious and decisive objection to this statement of Intrinsicness. Let *c* be the lighting of a fuse, and let *e* be an explosion some time later. Supposing *c* to be a cause of *e*, how could this fact possibly be *intrinsic* to the pair (*c*,*e*)? Surely the facts that make it the case that *c* causes *e* would need to include one that is *extrinsic* to the pair—namely, the fact that *this* fuse is connected to *that* bomb. So the claim that causation is a relation intrinsic to the events that exhibit it seems a non-starter.

But a non-reductionist can respond that the objection presupposes an ontology at odds with—because more spartan than—the ontology that he is committed to. As I understand it, the latter ontology simply *includes* causation as one of the fundamental properties and relations that characterize the world; the more spartan alternative does not. This contrast closely parallels another major difference of opinion that characterizes the literature on causation: namely, whether causal facts *reduce to* facts about what happens, together with facts about the fundamental laws of nature that govern what happens. One who says "yes" will, by way of clarifying her reductionist position, likely endorse something like the following picture: *All* facts about the world somehow reduce to or are constituted by or obtain in virtue of facts about which fundamental physical objects instantiate which fundamental physical properties and relations, together with facts about the fundamental physical laws.[6] And she will add that the list of

[5] For a representative statement, see Menzies 1999.

[6] Of course there is room for further dispute among reductionists: Some will say that to include facts about the laws is redundant, since these themselves reduce to facts about the instantiation of properties and relations by things. Some will insist that "physical objects" had better include spacetime itself. Some will demur about the restriction to the physical, wishing to allow, for example, that minds might be among the fundamental objects. (Note that such dualism does nothing at all to threaten the *reductionist* element in the physicalist position sketched in the text.) And there will be further disagreement about whether the reductionist claim is necessary or contingent—and if

fundamental relations is in all likelihood extremely *short*—perhaps comprising just spatiotemporal relations, perhaps a few more, e.g. if certain versions of quantum mechanics are right. At any rate, it does *not*, on her view, include causation itself. By contrast, her principal opponent thinks there is no hope of reducing causation to such a limited set of facts—and not for fancy reasons, but simply because he thinks (typically on the basis of various thought-experiments: see Tooley 1990) that causal facts don't even *supervene* on these other facts. He—at least, in some of his incarnations—therefore sees fit to introduce causation itself as a further fundamental relation. And it is therefore open to him to further stipulate that this relation is an intrinsic one.

I will side with the reductionist, although without making any attempt to resolve her dispute with the antireductionist. It is important to realize, however, that the appeal of Intrinsicness does not at all depend on endorsing, tacitly or otherwise, the kind of antireductionist view about causation just sketched. Think again about Suzy and Billy. When we judge that the extrinsic addition of Billy and his throw is irrelevant to the causal status of Suzy's throw, this judgment does *not* turn on conceiving of her throw as somehow directly connected to the shattering by the instantiation of a fundamental, metaphysically primitive causal relation. It turns, rather, on noticing at once that the addition does not affect in any relevant way the intrinsic character—the *ordinary*, garden-variety intrinsic character—of the process connecting Suzy's throw to the shattering.

So much for the first of the roads not taken. To map out the second road requires another brief digression. Distinguish two very different varieties of reductionism about causation. The first variety seeks to define some kind of *nomological entailment* relation that holds between cause-effect pairs. Examples will suffice to give the idea: We might hold that *c* is a cause of *e* iff, from the proposition that *c* occurs, together with the proposition that encapsulates the

contingent, contingent on *what*. None of these disagreements matter for present purposes; what is important is the philosophical sentiment that unites all the various forms of reductionism about causation.

fundamental laws, it follows (with metaphysical necessity, say) that *e* occurs. Or we might try something more sophisticated: *c* is a cause of *e* iff, from the proposition that *c* occurs, together with the proposition that encapsulates the fundamental laws, together with some suitably chosen proposition about conditions that obtain at the time of *c*'s occurrence, it follows that *e* occurs, where it is essential to the entailment that we include the premise that *c* occurs (Mackie 1965, more or less). Or we might hold that *c* is a cause of *e* iff, had *c* not occurred, *e* would not have occurred—where this entailment relation between the proposition that *c* does not occur and that *e* does not occur counts as nomological because of the central role played by the fundamental laws in fixing the truth-conditions of the counterfactual. Or we might hold that causation is the ancestral of this relation of counterfactual dependence (Lewis 1986b). Examples of this kind of approach abound in the literature, becoming more and more intricate the closer one gets to the present. As will become clear (§6), I endorse this approach (though not in its currently popular counterfactual form), and in fact will argue that one of the virtues of Intrinsicness is that it provides the key to making a nomological entailment account work, even in the face of stubborn cases like Suzy First.

Contrast a very different kind of reductionist account, which seeks to define causation in terms of some kind of *physical connection*—often involving the "transfer" of some physical quantity—between the cause and its effect. Again, I'll lean on examples: We might say, with Fair (1979), that causation consists in the transfer of *energy*. Or, with Ehring (1997), in the transfer of a *trope*. Or, with Dowe (1992) and Salmon (1994), in the transfer of some *conserved quantity*—where appeal is made to fundamental physics for an inventory of such quantities. This last example shows that fundamental laws have, or at least can have, a place in such accounts; all the same, their role is evidently much less direct than it is in nomological entailment accounts.

Physical connection accounts may vindicate some version of an Intrinsicness thesis—at least, insofar as the physical connections appealed to are intrinsic to processes that exhibit them.[7] In addition, some such account may well prove successful at capturing what causation is in worlds with laws like ours—*provided* we restrict our attention to causation between microphysical events. The qualifications appears unavoidable, as there seems little hope of employing the simple and austere connections of these accounts to map out causal relations in the messy macroscopic realm, or in possible worlds whose fundamental physics doesn't involve the transfer of anything. With respect to the former problem, consider that I drank coffee this afternoon, and a short while later became jittery as a result. It's not terribly plausible to think that we could arrive at some illuminating analysis of this causal process by focusing on the transfer of energy, momentum, charge, or any other fundamental physical quantity, or by searching for tropes that pass from the coffee to my body. After all, I drank *milk* with my coffee, and doing *that* presumably had nothing to do with my jitters—even though it affected the transfer of fundamental physical quantities to my body no less than did the coffee-drinking.[8] So I see no interest in trying to understand Intrinsicness by means of such theories, since their scope is far too limited.

I turn now to what I hope will prove a more fruitful way of developing the Intrinsicness thesis, taking a necessary detour through some further preliminaries.

§3 Assumptions

Let me now lay out some basic assumptions that will serve as my starting points. As noted—and like many philosophers who work on causation—I am going to assume that the *fundamental*

[7] Fair's position may be a counterexample, since *kinetic* energy, at least, has its value only relative to an inertial frame of reference.

[8] Granted, a detailed *scientific* understanding of the causal process leading from coffee-drinking to jitters would require examining its relevant microconsituents, and the relevant relations among them. But notice that use of the word "relevant" is inescapable here: e.g., the coffee transfers *heat* to my body, but that fact presumably does not matter to why the coffee makes me jittery. A proper philosophical account of causation ought to illuminate *what it is* about drinking the coffee that was causally responsible for my jitters, and it is a mark against transference theories that they lack the resources to do so. See Hall and Paul 2003 for more details.

causal structure of the world is given by the purely contingent facts about what happens, together with the facts about what the fundamental laws are that govern what happens. I think of these fundamental laws on the model of physics: namely, as something like rules that specify how complete physical states of the world generate successive physical states.[9] If you like, then, the fundamental causal facts take the following form: complete-physical-state-P_1 causes complete-physical-state-P_2. There is, of course, ample room, and even a pressing need, within such a conception for much more circumscribed causal facts involving correspondingly circumscribed and localized events; e.g., the fact that my pressing this letter on the keyboard caused that letter to appear on the screen. But again, like many philosophers, I take the further step of holding that these causal facts somehow *reduce to* facts about what happens, together with the laws that govern what happens—and do so in a way that careful philosophical analysis can elucidate. By "reduce to" I mean something stronger than merely "supervene on"; for the former relation is patently asymmetric, whereas the latter is not. (And in the present case, it would not be that surprising if it turned out that both the facts about what happens and the facts about the fundamental laws supervene on the totality of causal facts expressible in the form "event *c* causes event *e*".) I will say a bit more later about the form that I think such a reductionist analysis should take.

I have implicitly suggested, and hereby explicitly endorse, the view that the basic causal relata are events. I won't try to say what "basic" means here, save to point out that some such qualification is necessary, since, after all, people, rocks, and other ordinary particulars can in some sense cause things, as perhaps can items from still other ontological categories (*facts*, as it might be). I also won't try to say what events are, although I will assume this much about them: like other particulars, they have detailed intrinsic natures. Such natures correspond to easily

[9] See Maudlin, 1996, for an excellent exposition of this conception of law. Even though this is the way I think of the laws, all that is really necessary for me to assume is that there is a firm distinction between worlds that are, and worlds that are not, *nomologically possible*. So it is compatible with everything I have to say that the world be infinitely complex, in the sense that it has no "fundamental" level (see Schaffer REF).

recognizable facts about events. Suppose, for example, that I throw a rock. Then there are facts about exactly how long the throw lasts, exactly what shape region it occupies at each moment of its occurrence, the exact mass the rock has at each of these moments, etc. I take these to be facts about the *intrinsic properties* of my throw. More generally, I assume that for any event *e*, the totality of such facts serve to specify exactly—but do no more than serve to specify exactly—how *e* occurs, down to the minutest detail. I will therefore further assume that it makes sense to compare any two events with respect to their intrinsic *similarity*. I will likewise assume that structures built up out of events[10] have intrinsic natures—exhausted, I will suppose, by the intrinsic natures of their constituent events, together with the facts about how those events are spatiotemporally juxtaposed with one another. (I count the time order of its constituent events as an intrinsic feature of an event structure, so that two such structures that are temporally inverted with respect to each other qualify, for that reason, as being intrinsically *dis*similar.) All of this will be important in what follows, for we will see that stating the Intrinsicness thesis carefully requires us to make use of the notion of intrinsic similarity between entire event structures.

I am going to make the simplifying assumption that the fundamental laws are deterministic, and that they permit neither backwards causation nor action at a temporal distance. At various points in what follows I will briefly indicate where complications arise when we give up these assumptions, but I won't pursue those complications in any detail.

Turn now to the Intrinsicness thesis. The crude statement of the thesis was that the causal structure of a process is intrinsic to it. We can arrive at a better statement if we focus on simple examples. I will choose the ones we started with—Suzy's throw, with and without Billy's backup throw—but represent their structure abstractly, by means of the "neuron diagrams" popularized by Lewis. Thus, in figure 1, neuron **a** fires, sending a stimulatory signal to **e**, causing

[10] "Built up" how? By mereological fusion, I suppose—although those uncomfortable with extending mereology into the domain of events could probably substitute some set-theoretical surrogate.

it to fire; we represent the firings of **a** and **e** by shading their respective circles, and represent the passage of the stimulatory signal by the arrow connecting these circles:

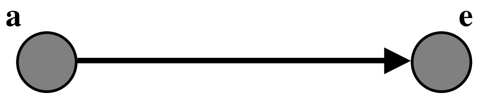

Figure 1

Note that the order of events is left to right.

In figure 2, exactly the same events happen, plus some additional ones: neuron **b** fires a few moments after **a**, sending a stimulatory signal to **e** that has almost reached **e** by the time **e** fires. (We represent this fact by drawing an arrow from **b** that doesn't quite reach **e**.)

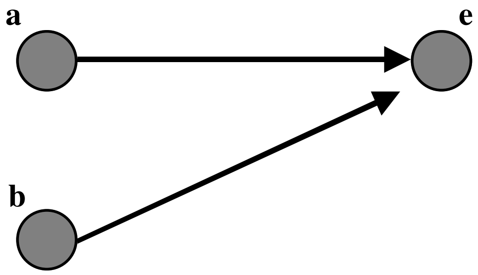

Figure 2

The addition of **b** and its signal does nothing at all to alter the status of *a* as a cause of *e*. (Throughout, I will use italicized letters to represent the events consisting in the firings of the associated neurons.) Here, then, is the thought: Fix on an event *e*, and take the structure consisting of that event together with all of its causes back to some earlier time t. Then altering the *environment* of that structure while holding the structure itself fixed cannot alter the fact that the events in the structure (distinct from *e*) are *causes* of *e*. In that sense, causal relations are *stable* under perturbations of the environment of the process exhibiting them.

That statement of the thesis has the distinct advantage of being both intuitive and plausible. But it is important to restate the thesis in a way that makes it clear what we mean when we talk about holding a structure of events "fixed" while changing the environment. Hence this final version of the thesis:

<u>Intrinsicness</u>: Let S be a structure of events consisting of event *e*, together with all of its causes back to some earlier time t. Let S' be a structure of events that intrinsically matches S in relevant respects, and that exists in a world with the same laws. Let *e*' be the event in S' that corresponds to *e* in S. Let *c* be some event in S distinct from *e*, and let *c*' be the event in S' that corresponds to *c*. Then *c*' is a cause of *e*'.

So stated, the thesis raises a number of questions. Five stand out.

§4 Questions and brief answers

<u>First question</u>: What is the motivation for the thesis?

<u>Answer</u>: First, there is the intuitive motivation brought out by contrasting cases like Suzy Alone and Suzy First. Second, the thesis offers an elegant solution—and perhaps the *only* solution—to what I am calling the "overdetermination problem" (§§5 and 7). Third, reflection on the *function* of our causal concepts suggests that at least one such function, and a central one, is to provide us with tools for the piecemeal representation of the nomological structure of our world; the <u>Intrinsicness</u> thesis, if correct, goes a long way towards explaining how our causal concepts can play such a role (§9).

<u>Second question</u>: Why is the thesis so complicated? In particular, why must S be so comprehensive as to include *all* the causes of *e* (back to some earlier time t)?

<u>Answer</u>: Causes produce their effects *by way of* causal intermediates, and *with the help of* other causes with which they combine; state the thesis in a way that ignores these truisms, and you expose it to trivial counterexamples. (§6)

Third question: Why the qualification "in relevant respects"?

Answer: Because non-causes can make small and irrelevant differences to the manner in which causes occur. Add Billy's throw so as to convert Suzy Alone into Suzy First, and you inevitably introduce ever-so-subtle changes in the flight path of her rock (due, for example, to gravitational attraction from Billy's rock). (§8)

Fourth question: What does this qualification mean?

Answer: Roughly this: S' must be intrinsically similar to S in those respects in which S and other "blueprints" like it are similar to each other. Begin with Suzy Alone, and introduce slight variations in her throw (or in the window) that make no difference to the causal structure of the processes that lead up to the shattering: the multitude of ways of doing so provides us with a set of "blueprints". These blueprints, in turn, allow us to set up standards of comparison by means of which we can distinguish relevant from irrelevant respects of similarity. (§8)

Fifth question: What are the consequences of the thesis?

Answer: It provides a solution to the overdetermination problem. It provides a novel argument that counterfactual dependence is not sufficient for causation, and likewise that there is no causation by omission. It provides the materials for a clean account of symmetric overdetermination, and of at least one notion of "causal efficacy". (§§7 and 8)

On to more detailed answers.

§5 The overdetermination problem

Let's start with reasons for believing the Intrinsicness thesis. I have already talked about the intuitive motivation for the thesis by comparing and contrasting the events depicted in figures 1 and 2. I now want to put those examples to use for a different purpose, in order to bring out the serious problem that the Intrinsicness thesis may offer the only good hope of solving.

Recall one of our starting points: I assumed that philosophical analysis should be able to show how causal facts reduce to facts about what happens, together with facts about what the laws are that govern what happens. As noted in §2, a thriving tradition has it that a successful analysis of causation should reveal that relation to be a kind of *nomological entailment* relation. Once again, some illustrative examples (of which the literature offers countless variations):

Crude sufficient condition account: *c* causes *e* iff *c* and *e* both occur, and from the fact that *c* occurs, together with the laws, it follows that *e* occurs. More precisely: *c* causes *e* in world w iff *c* and *e* both occur in w, and in any world w' in which *c* occurs, and which has the same laws as w, *e* occurs.

Crude necessary condition account: *c* causes *e* iff *c* and *e* both occur, and from the fact that *e* occurs, together with the laws, it follows that *c* occurs. More precisely: *c* causes *e* in world w iff *c* and *e* both occur in w, and in any world w' in which *e* occurs, and which has the same laws as w, *c* occurs.

Mackie-style regularity account: *c* causes *e* iff *c* and *e* both occur, and from the fact that *c* occurs, together with some suitable auxiliary premises describing contingent facts about the circumstances in which *c* occurs, together with the laws, it follows that *e* occurs; but this fact does *not* follow from the auxiliary premises and the laws alone. (This is roughly Mackie's view: causes are necessary parts of sufficient conditions for their effects.)

Simple counterfactual account: *c* causes *e* iff *c* and *e* both occur, and had *c* not occurred, *e* would not have occurred.

Lewis-style counterfactual account (circa 1976): *c* causes *e* iff *c* and *e* both occur, and there is a (possibly empty) set of events $\{d_1, d_2,..., d_n\}$ such that if *c* had not occurred, d_1 would not have occurred; and if d_1 had not occurred, d_2 would not have occurred;... and if d_n had not occurred, *e* would not have occurred. (This is the analysis offered in the classic Lewis 1976.)

Updated Lewis account (circa 2000): *c* causes *e* iff *c* and *e* both occur, and there is a (possibly empty) set of events $\{d_1, d_2,..., d_n\}$ such that *c* influences d_1, d_1 influences d_2,..., d_n influences *e*; one event *influences* another just in case, roughly, any of a large number of slight counterfactual variations in the first would be followed by corresponding variations in the second. (see Lewis 2000)

In each case, the analysis attempts to show causation to be a kind of *entailment* relation (using that term in a permissive sense, so that it covers, e.g., counterfactual accounts as well as regularity accounts) between the fact that the cause occurs and the fact that the effect occurs (sometimes going from cause to effect, sometimes from effect to cause), where that relation is, crucially, mediated by the fundamental laws. While the search for such an entailment relation can easily go astray—e.g., none of the foregoing analyses succeeds—the motivation behind it is, to my mind, fundamentally sound. At the very least, the project of pursuing this kind of reductionist analysis of causation is fruitful enough, even where it fails, to deserve serious and sustained attention.

Reflecting on figure 2, however, it can seem utterly hopeless. For what is so striking and dispiriting about this example is that the events *a* and *b* bear, apparently, *exactly the same* nomological entailment relations to *e*. Each is a necessary part of some sufficient condition for *e*; *e* counterfactually depends on neither—but would so depend, had the other not occurred; both *a* and *b* influence *e* (indeed, to pretty much the same degree); etc. Moreover, the usual tricks for handling cases of overdetermination—for example, interpolating some intermediate event such that the cause bears the appropriate entailment relation to it and it bears that relation to the effect (see Lewis 1986b)—seem to be of no avail.[11] Notice, in this regard, that the close parallel in nomological entailment relations to the effect persists as we move from *a* and *b* to successive pairs of events more and more causally proximate to the effect.

At this point, we might want to try more desperate measures. For example, perhaps it matters crucially that if **a** had not fired, **e** would have fired at a slightly different time, whereas the same would not be true if **b** had not fired. More generally, we might want to distinguish causes by saying that they are the events upon which the detailed manner of occurrence of the effect depends. But not only will this maneuver let in far too many events as causes (namely, by

[11] Terminological point: I am calling cases like figure 2 cases of "overdetermination", where some (e.g. Lewis 1986b) would use the term "redundant causation"—reserving "overdetermination" for *symmetric* cases (as when two neurons simultaneously stimulate a third).

erasing the distinction between events that *cause* a given effect and events that merely make a difference to the *way* in which that effect occurs), it is also embarrassingly easy to tinker with our example so as to render it ineffective. Suppose, for example, that the signal from **a** exerts a retarding force on the signal from **b**, slowing it down slightly. Suppose that the exact strength of this force is such that, had **a** not fired, the signal from **b** would have arrived at **e** at exactly the same time that the signal from **a** in fact arrives. Then not only does the firing of **e** not depend on the firing of **a**, but nothing *about* the firing of **e**—not its timing, nor any other feature of its manner of occurrence—depends on the firing of **a**. (We can further suppose that the retarding force is of just the right strength that, for exactly the same reason, *e* exhibits a total lack of dependence on each of the events that consist in the passage of the stimulatory signal from **a** to **e**.)

More desperate measures still? Perhaps we could insist that causes be connected to their effects via spatiotemporally continuous causal chains. But of course there *are* spatiotemporally continuous chains connecting *b* to *e*. For example, the stimulatory signal from **b** emits radiation as it moves along, and some of this radiation strikes **e** at the exact moment that it begins to fire. And anyway, it seems drastic to be forced to accommodate such a simple example by ruling out action at a distance a priori. (Note that we would be ruling out not merely action at a *temporal* distance—a restriction which we need not find so troubling—but also action at a spatial distance. And that is the one kind of action at a distance for which the history of physics provides some precedent.)

There are, of course, other and still more elaborate constructions that one might try. Vast tracts of the literature on causation are littered with their remnants. But there is no further profit to be had in sifting through them.[12] Enough has been said to underscore how serious is the problem posed by such simple cases of overdetermination as that depicted in figure 2. I will just add that their very simplicity argues for a correspondingly simple solution. It would be

[12] For excruciatingly detailed discussion, see Hall & Paul 2002.

disappointing if we could only handle such mundane examples by means of an intricate technical apparatus. We will see, happily, that the Intrinsicness thesis offers an attractively elegant solution. The key idea is to find a nomological entailment relation that succeeds in capturing the causal structure of a process when circumstances are *nice* (as they are, for example, in figure 1), and to use the Intrinsicness thesis to "transfer" that causal structure into circumstances that are not so nice (e.g., figure 2). But before developing this idea, let us first consider why that thesis needs to have the form I have given it.

§6 Motivating the constraints

What the Intrinsicness thesis allows us to do is to fix on one structure of events, and then to "transfer" its causal characteristics over to another, intrinsically matching structure. But the two structures must exist in worlds with the same laws, and the structure we begin with must be comprehensive enough to consist of an event together with all of its causes back to some earlier time. It may be unclear why these constraints should be in place. That is what I will explain in this section. What I will do is to show by example that the most obvious ways of relaxing them do not give us a workable version of the Intrinsicness thesis.

§6.1 Laws

Begin with the constraint on laws. Recall figure 1: **a** fires, emitting a stimulatory signal which travels to **e**, at which point **e** fires. The relevant underlying laws hold that **e** will fire iff (and when) a stimulatory signal reaches it. But suppose the laws to be radically different. Suppose they dictate that **e** fires spontaneously every two seconds. (By "spontaneously", I do not mean *without any causes*, but just without any *external* causes.) Suppose further that the laws specify no role for the signals that neurons sometimes emit when they fire. Such laws permit an exact duplicate of the events depicted in figure 1: **e** has been firing spontaneously, every two seconds, since the dawn of time; **a** fires, emitting its signal in the direction of **e**; that signal reaches **e**—purely by coincidence—at exactly the moment at which **e** is beginning one of its firings. To be sure, in realistic examples we could probably discern subtle intrinsic differences in the two event structures, differences that bore witness to the drastic difference in underlying laws. But I would have you imagine that in our highly abstracted neuron worlds, the *only* difference is in the underlying laws. The *a-e* event structures, in each case, are intrinsically *exactly* alike.

The *causal* structures are completely different, of course. The appropriate conclusion is one that might seem a truism: the fundamental laws play an essential role in fixing causal structure. The best way to implement this lesson in the development of the Intrinsicness thesis is to require that intrinsic similarity make for similarity in causal structure, *provided* that these laws are held fixed.[13]

On to the less obvious constraint that requires us to pick an event structure that consists of an event *e*, together with *all* of its causes back to some earlier time t. I will explain in stages why this constraint should be in place.

§6.2 Causal mediation

Suppose we tried to defend this simpler statement of the Intrinsicness thesis: When a pair of events (*c*, *e*) intrinsically matches another pair of events (*c*', *e*'), and both pairs exist in worlds with the same laws, then the same causal relations obtain in each case—i.e., *c* is a cause of *e* (likewise *e* of *c*) iff *c*' is a cause of *e*' (respectively, *e*' of *c*'). That is hopeless, even when we make clear that in order for such pairs to intrinsically match each other, they must duplicate the spatiotemporal relations exhibited by their members. For whether we have causation in each case obviously depends on what sorts of connections, if any, there *are* between the events in question. And the existence or lack thereof of such connections is not going to be a fact intrinsic to the pairs.

Consider figure 3:

[13] One might think this proviso unnecessary, on the grounds that the fact that a process is governed by such-and-such laws is itself an *intrinsic feature* of that process. As was implicit in my discussion of the spontaneously firing neuron, I disagree. But nothing hangs on this question; if I am wrong, the worst that happens is that my statement of the Intrinsicness thesis is somewhat redundant.

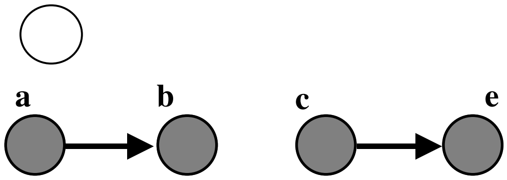

Figure 3

Here, **a** fires, sending a signal to **b**, causing it to fire. A little later, **c** fires (as a result of causes that are not depicted—but *not* as a result of *b*), sending a signal to **e**, causing it to fire. The (*a, e*) pair depicted in figure 3 is, we can suppose, intrinsically *exactly* like the (*a, e*) pair depicted in figure 1. But the causal relation that obtains in this latter case does not transfer to the former.

There are two distinct lessons to learn from this case. For the first, suppose that we have some event structure, picked out in some manner, that consists in part of event *a* and later event *e*. Suppose further that *a* is *not* a cause of *e*. This structure might, for example, be the one consisting of all the events depicted in figure 3. Then it will almost always be possible to embed an intrinsic duplicate of this structure in an environment that provides sufficient "connecting material" to make the event corresponding to *a* a cause of the event corresponding to *e*—as for example in figure 4, where we insert a stimulatory connection between **b** and **c**:

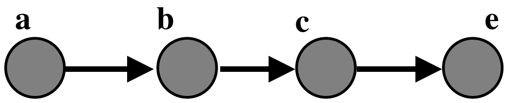

Figure 4

Granted, it might in rare cases *not* be possible, namely if the structure with which we begin is so "fat" that it leaves no room for such connecting material. But there is no interest in trying to come up with a condition that will guarantee such "fatness". So the first lesson is this: In constructing our Intrinsicness thesis, we should expect to come up with a principle that only enables us to transfer *positive* causal characteristics from one event structure to another. Once we have identified our “blueprint” structure S, and have picked out some event structure S’ that intrinsically matches S, we should expect only that where *c* in S causes *e* in S, the intrinsic similarity between S and S’ guarantees that *c*’ in S’ causes *e*’ in S’ (where, as usual, *c*’ is the event in S’ corresponding to *c* in S, etc.). We should *not* expect, in addition, that where *c* in S does *not* cause *e* in S, the intrinsic similarity between S and S’ guarantees that *c*’ in S’ does not cause *e*’ in S’.

The second lesson concerns the way in which the event structure we begin with must be specified. For suppose this structure contains events *c* and *e*, where *c* is a cause of *e*; but suppose it *fails* to contain every event that is causally intermediate between *c* and *e*—that is, every event that is both an effect of *c* and a cause of *e*. Then, as figures 1 and 3 show, it will almost always be possible to embed an intrinsic duplicate of this structure in an environment that “deletes” the relevant intermediate causes, and that guarantees that *e* is brought about by other, *independent* means. In any such environment, *c* (or its duplicate) will no longer be a cause of (the duplicate of) *e*. The simplest way to avoid this problem is to require that the event structure with which we begin include, for every pair of events *c* and *e* where *c* is a cause of *e*, all the causal intermediates between *c* and *e*.

§6.3 Causal combination

So far, we have come up with constraints motivated by the need to attend to the essential role that laws play in fixing causal structure, and to the fact that causes typically yield their effects only *by way of* causal intermediates. But causes also typically yield their effects only with the *help* of other causes with which they combine, and this fact introduces the need for one more constraint, as a pair of examples will show.

Suppose that neurons can fire with different “polarities”—say, positive and negative—emitting stimulatory signals with those same polarities. Suppose that neuron **e** in figure 5 requires two stimulatory signals in order to fire (we represent this fact by drawing it with a thick border), and further requires that these signals be of the same polarity. Suppose finally that the way in which **e** fires is not at all sensitive to whether it receives two positively polarized or two negatively polarized stimulatory signals:

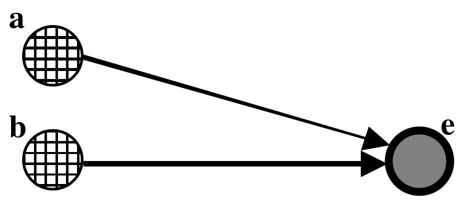

Figure 5

Here, **a** and **b** both fire with positive polarity (represented by hatching their circles), sending positively polarized signals to **e**, which fires as a result. Now consider figure 6:

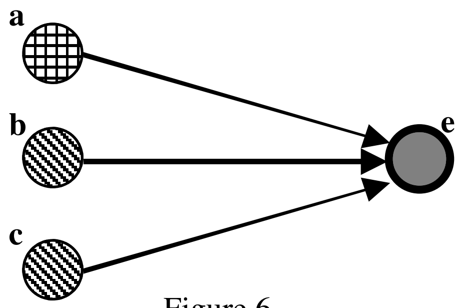

Figure 6

Again, **a** fires with positive polarity, sending a positively polarized signal to **e**. But **e** receives only one such signal; the reason it fires is that it also receives two negatively polarized signals (represented by diagonal lines), one from **b** and one from **c**. It is sheer coincidence that these signals reach it, and thus cause it to fire, at exactly the same moment that the signal from **a** reaches it.

In figure 5, *a* is a cause of *e*, but in figure 6, *a* is not a cause of *e*. Nevertheless, if we pick out, from the events depicted in figure 5, a structure consisting of *a*, *e*, and all the events causally intermediate between *a* and *e*, we see this structure duplicated *exactly* in the events depicted in figure 6. We are obviously not entitled to use this similarity as grounds for transferring the causal characteristics of the *a-e* sequence in figure 5 over to the *a-e* sequence in figure 6.

The problem is that we have not properly incorporated the fact that in figure 5, *a* does not cause *e* all by itself: it requires the assistance, as a kind of combining cause, of *b*. We should therefore require that when the event structure S contains *c* and *e*, with *c* a cause of *e*, then it must also contain the other causes with which *c* combines to bring about *e* (typically, these will be those causes of *e* that are contemporaneous with *c*).

Let us now put all this together. We set out to provide a recipe for specifying an event structure in such a way as to guarantee that any other structure that intrinsically matches it will also share its causal characteristics. We have seen that the two structures need to exist in worlds with the same laws. We have seen that we can only expect sharing of "positive" causal characteristics. We have seen that the event structure we begin with must meet a certain closure condition: it must be comprehensive enough that when it includes an event *e* and one of its causes *c*, it likewise includes all events causally intermediate between *c* and *e*, as well as the events with which *c* combines to bring about *e*. The next step is to replace these two explicitly causal notions—"causal intermediate" and "combining cause"—with non-causal surrogates. What allows us to do so is the assumption that there is neither action at a temporal distance nor backwards causation. Without the first restriction, we would have to contend with the possibility that a cause *c* combines with other causes not contemporaneous with it. (Example: Two magicians cast spells, at different times, that jointly act at a temporal distance to bring about some effect.) Without the second, we would have to contend with the possibility that, even where *c* precedes its effect *e*, some of the events causally intermediate between *c* and *e* lie outside the time interval between them. Such possibilities make trouble for a clean statement of the Intrinsicness thesis, as well as for the project of using the Intrinsicness thesis to extend the scope of a reductive analysis of causation (see the next section); it would be distracting and profitless, at this point, to explore how to handle this trouble. Instead, we will take advantage of our assumptions to arrive at a particularly simple specification of the kind of event structure that meets our closure condition: it should consist of an event *e*, together with all of its causes back to some earlier time t. It is therefore such structures—structures that we will henceforth call "*e*-blueprints"—that our statement of the Intrinsicness thesis concerns.

§7 Consequences of the thesis

Let us turn to three applications of the Intrinsicness thesis. The first is a solution to the problem of overdetermination (which yields, as a kind of byproduct, an account of *symmetric* overdetermination as well). The second is an argument that counterfactual dependence does not suffice for causation. The third is an argument that there is no causation by omission. I will take these in order.

§7.1 Solving the problem of overdetermination

Recall the contrast between the two cases Suzy Alone and Suzy First. Here is a natural way to think about that contrast: We can get to Suzy First by starting with Suzy Alone, and simply adding things to the environment—Billy, his throw, the flight of his rock through the air, etc. The crucial point is that while making these additions, we can *hold fixed* the event structure consisting of Suzy's throw, the flight of her rock to the window, the shattering of the window, and, for completeness, the prior presence of the window itself. That is, we can hold fixed a structure consisting of the shattering of the window together with all of its causes back to the time of Suzy's throw. And what I mean by "hold fixed" is that making these additions to the environment does not alter—or at least, does not alter in any relevant respect—the intrinsic character of that structure. Since Suzy's throw in Suzy Alone is, obviously, a cause of the shattering, it therefore follows from the Intrinsicness thesis that in Suzy First her throw is a cause of the shattering. For short, Suzy First contains within it an event structure that (i) includes Suzy's throw as part; and (ii) intrinsically matches some shattering-blueprint (namely, the one provided by Suzy Alone). It is in virtue of (i) and (ii) that Suzy's throw counts as a cause of the shattering.

Now try to do the same with Billy's throw. We might start by imagining a situation—call it "Billy Alone"—in which Billy throws a rock at the window, breaking it, and Suzy is absent. We might then try to imagine ringing changes on this situation so as to add Suzy and her throw, arriving by these changes at Suzy First. But notice that if we begin, in Billy Alone, with a structure consisting of the shattering together with all of its causes back to the time of Billy's throw, then we cannot possibly hold this structure fixed in intrinsic respects while morphing the situation into Suzy First. In particular, the spatiotemporal relations that the shattering bears to the events constituting the flight of Billy's rock through the air will change in crucial respects: a spatiotemporal gap will open up where formerly there was none.

That is the intuitive idea behind the way in which the Intrinsicness thesis allows us to distinguish the causal status of Suzy's throw from the status of Billy's throw, in Suzy First. It is important that its intuitive force not be obscured by the following unavoidable caveat: namely, that what matters is not merely that in morphing Billy Alone into Suzy First we are forced to make *changes* to the intrinsic character of the given event structure (for the same is true, when we change Suzy Alone into Suzy First), but that these changes must inevitably be changes in *relevant respects*. I sketch, in the next section, an account of what distinguishes relevant from irrelevant respects. For now, I want to show how the Intrinsicness thesis can be welded to a reductive analysis of causation in such a way as to enable that analysis to overcome the problem of overdetermination.

Suppose we have an analysis of causation that is *partial*, in that it issues verdicts about some situations, but simply falls silent about others. Suppose further that when it issues a verdict, it never makes a mistake, and that in at least some situations it does more than merely to identify the causes of a given event: in addition, the analysis successfully identifies them as being *all* the causes. In particular, suppose that the analysis, as applied to Suzy Alone, correctly identifies all the causes of the shattering (back to the time of Suzy's throw), and correctly identifies them *as being* all the causes. But, finally, suppose the analysis fall silent about the more complicated Suzy First.

It will help to have an example, so let me quickly sketch an analysis with these features. I do not mean to argue for this analysis here (though something like it is developed and defended in my 2003 and 2004), but only to use it for purposes of illustration:

Consider an event *e*, and some time t earlier than the time at which *e* occurs. Let the set S consist of events occurring at t. Say that S is *sufficient* for *e* if and only if *e* would still have occurred, had only the events in S occurred at t. Say that S is *minimally sufficient* for *e* if and only if it is sufficient for *e*, but no proper subset of it is. Then the analysis says that *if* (note: not "if and only if") S is the unique set of t-events minimally sufficient for *e*, then the events in S are all the *causes* of *e* occurring at t. Observe that this analysis falls squarely within the tradition of nomological entailment accounts.

In Suzy Alone, we can be sure that at each time before the shattering, there is a unique set of events minimally sufficient for the shattering. The analysis will therefore say that the events in such sets are all causes of the shattering, and indeed that they are the *only* causes (occurring at the given time). The problem with Suzy First, of course, is that at each time before the shattering, there is more than one set of events minimally sufficient for the shattering: on the one hand, there is a set containing (depending on the time) either Suzy's throw or one of the events in the flight of her rock to the window; on the other hand, there is a set containing either *Billy's* throw or one of the events in the flight of *his* rock to the window. Since the analysis only aims at providing a sufficient condition for causation—a condition that *requires* a unique minimally sufficient set—it falls silent about Suzy First. Thus, while the analysis succeeds in identifying the complete causal history of the shattering in Suzy Alone, it says nothing about Suzy First.

But *any* analysis that succeeds in this former task can automatically be extended to cover Suzy First by the addition of the Intrinsicness thesis.[14] For it is a consequence of the analysis that, in Suzy Alone, the structure of events consisting of the shattering, together with Suzy's

[14] Provided, that is, that the analysis is *consistent* with the Intrinsicness thesis. We'll see shortly how important this qualification is.

throw and the flight of her rock, together with the prior presence of the window itself, just *is* a shattering-blueprint: a structure consisting of the shattering together with all of its causes back to the time of Suzy's throw. Since Suzy First contains an intrinsic duplicate of this structure, it follows from the Intrinsicness thesis that in Suzy First, Suzy's throw is a cause of the shattering.

The picture is this: We started by observing that simple cases of asymmetric overdetermination threatened to block the progress of any analysis of causation that attempts to reduce causation to some kind of nomological entailment relation. We therefore retreat, attempting to come up with an analysis that *doesn't get those cases wrong*—avoiding such refutation simply by falling silent about them. We then see that as long as our analysis can get the "blueprint" cases—cases like Suzy Alone—right, then, augmented by the Intrinsicness thesis, it will automatically cover the troublesome cases of asymmetric overdetermination as well. It is in this way that the Intrinsicness thesis dissolves the threat to the project of analyzing causation posed by asymmetric overdetermination.

At least, this will be so *provided* that we can arrive at any case of asymmetric overdetermination by starting with some blueprint case, and adding details to it extrinsic to the causal process featured in it (in the way that we could arrive at Suzy First by starting with Suzy Alone). It is plausible that we will always be able to do so. After all, what marks a case of asymmetric overdetermination *as* such? Presumably, that it contains, in addition to the genuine causal processes that yield the given effect, backup processes that would have done so, had the genuine causes been absent. Working backwards, then, we can construct a blueprint for such a case simply by stripping away these idle backup processes. The only worry would be that in doing so, we might inevitably—i.e., with nomological necessity—substantially alter the intrinsic character of the structure consisting of the effect together with the genuine causal processes. But if the absence of the "idle" backup processes would *have* to make such a difference, then it is no longer clear that they are really idle: Some of them, at least, should have been counted among the genuine causes in the first place.

Before proceeding to the more destructive consequences of the Intrinsicness thesis, let us pause to consider how it lends itself to an account of *symmetric* causal overdetermination. Full discussion of this topic would take us too far afield, so I will limit myself to a single illustrative example. Consider

Same Time: Billy and Suzy both throw rocks at a window, each with sufficient force to shatter it. The rocks strike the window at exactly the same time. The window breaks.

This kind of case puzzles us. Should Billy's throw count as a cause of the shattering? If so, then so should Suzy's; but isn't that redundant? If not, then neither should Suzy's; but doesn't that leave the shattering uncaused? Never mind. It is not our concern here to resolve these puzzles, but rather to answer a prior question: What distinguishes cases like this *as* cases of symmetric causal overdetermination—the kinds of cases that give rise to these puzzles?

The literature provides various answers, none particularly successful.[15] It has not provided the one that I think is most attractive, and that I will now sketch. Consider two ways that we could arrive at Same Time. First, we could begin with Suzy Alone, and then add stuff to the environment: Billy, his throw, the flight of his rock to the window. Second, we could begin with Billy Alone, and add stuff to the environment: Suzy, her throw, the flight of her rock to the window. In each case, the extrinsic additions do not alter in any relevant respect the intrinsic characteristics of the event structures we start with.[16] So it follows from the Intrinsicness thesis that in Same Time, *both* Billy's and Suzy's throws are causes of the shattering. More to the point, it follows that Same Time contains within it two distinct event structures, each of which intrinsically matches some blueprint: one of these event structures matches the blueprint provided by Suzy Alone, the other the blueprint provided by Billy Alone. In this sense, then, Same Time contains within it two distinct *complete causal histories* of the shattering—where

[15] For an extended argument that this is the case, see my 2004.

[16] Or if they do—say, by dramatically altering the manner in which the window shatters—then we should reverse our original verdict, and count Same Time a case of *joint* causation rather than a case of symmetric overdetermination.

what marks an event structure as a complete causal history of some event *e* (back to some time t) is that, when augmented by *e* itself, the resulting structure intrinsically matches some *e*-blueprint. I suggest that in general, symmetrically overdetermined events are exactly those with multiple complete causal histories.

§7.2 Intrinsicness and the counterfactual analysis

It might seem that this *blueprint strategy* for handling asymmetric overdetermination could be pursued in any of a number of ways, by grafting the Intrinsicness thesis onto any of a number of partial analyses of causation. In particular, it might seem that it could be grafted onto a partial *counterfactual* analysis of causation. Suppose, for example, that one started with the simplest counterfactual analysis, but took it to provide only a sufficient condition for causation: *c* is a cause of *e* if *e* would not have occurred if *c* had not occurred. And suppose further that one could somehow add a "completeness" condition allowing one to tell, at least for sufficiently "nice" situations, that the sufficient condition has identified *all* the causes of some given event. Add the Intrinsicness thesis, and it now appears that the counterfactual analyst has a solution to the problem of overdetermination.

Not quite. We can use the Intrinsicness thesis to extend a partial analysis of causation only if that thesis is *consistent* with the analysis with which we begin. The problem for the counterfactual analyst is that Intrinsicness contradicts the claim that counterfactual dependence suffices for causation. More carefully, the combination of these two claims yields consequences so dramatically at odds with our basic intuitions about causation as to make it implausible that an analysis founded jointly upon them will have any utility.

To see why, consider a situation in which *e* depends on *c* only because *c* cancels some threat to *e*:

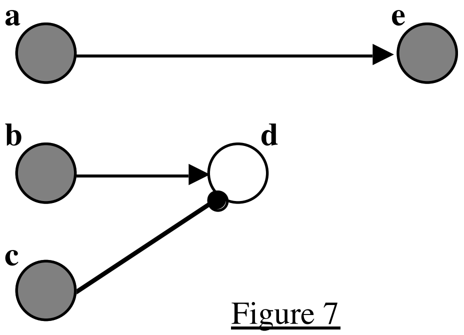

Figure 7

In figure 7, neuron **c** fires at time 0, sending an inhibitory signal to **d** that switches it off at time 1. (Lines with blobs at the end represent these inhibitory signals.) At time 2, **b** fires, emitting a stimulatory signal that reaches **d** at time 3 (and that has no effect). Also at time 3, **a** fires, sending a stimulatory signal to **e** which causes it to fire at time 5. Figure 8 depicts what would have happened had **c** not fired:

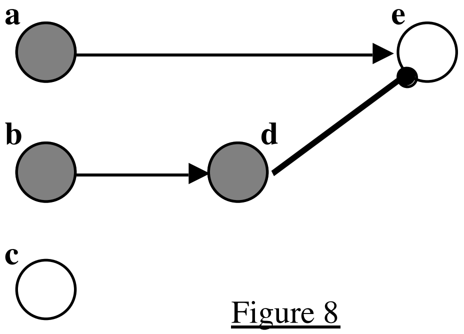

Figure 8

Here, the stimulatory signal from **b** causes **d** to fire; **d** in turn sends an inhibitory signal to **e** that reaches it at time 4, turning it off before the signal from **a** arrives. Consequently, **e** never fires.

These diagrams display as clean a case of counterfactual dependence as could be desired: In figure 7 **c** fires; **e** fires; but if **c** had not fired, **e** would not have fired. Suppose we conclude that *e* has among its causes *c*, along with the events making up the passage of the inhibitory signal from **c** to **d**, and the protracted event consisting in **d**'s being in a turned off state from time 1 to time 3. That is, we collect together as causes of *e* not merely *a*, together with the events making up the passage of the stimulatory signal from **a** to **e**, but also all other events upon which *e* depends. If the core thesis behind the counterfactual analysis—that counterfactual dependence suffices for causation—is correct, then the structure consisting of *e*, together with all of these events, *just is* a structure consisting of *e*, together with all of its causes back to time 0.

The problem is that this structure will be duplicated—duplicated *exactly*, down to the most microscopic detail—in a situation in which *c* is clearly not a cause of *e*, simply because there was no threat to be canceled. For suppose that **b** never fires. More dramatically, suppose that **b** simply does not exist, and that there is nothing taking its place that threatens to stimulate **d**. Then we would have the situation depicted in figure 9:

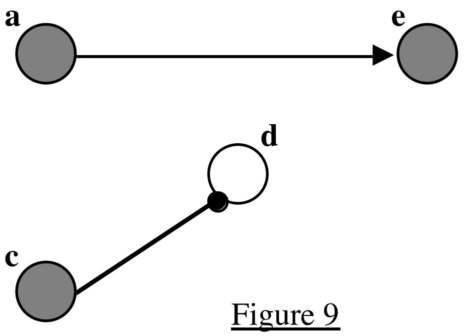

Figure 9

c fires, sending an inhibitory signal to **d**, switching it off. Meanwhile, **a** fires, sending a signal to **e**, causing it to fire. Here, *e* clearly does not depend on *c*. More to the point, *c* can in no sense be considered causally responsible for *e*. In figure 7, its causal responsibility was exhausted by the fact that it canceled a threat to *e*; figure 9, containing no threat to be canceled, does not allows *c* even that limited role.

There is, to be sure, a sense in which *c* can be said to “safeguard” *e*: for *c* guarantees that certain threats, *were* they to materialize, *would* be canceled. Far from helping the counterfactual analyst, this point helps bring her mistake into sharper focus, for there seems little future in an approach to causation that cannot distinguish between safeguarding and causing. The crucial point is that when something threatens to prevent *e*, the threat’s status *as* a threat will typically result partly from factors *extrinsic* to the causal history of *e*. Once you see the trick, it is child’s play to multiply examples indefinitely. Partly just for fun, but partly also to return from the neuron world to the real world, I will provide one more.

Disappointed: Suzy and Billy want to have lunch together. Suzy leaves a note pinned to her front door, telling Billy where to meet her. On his way to her house, Billy trips over a tree root. He tries, but fails, to catch himself in time. He falls over. He bumps his head, and is knocked cold. He never reaches Suzy's house, and so never sees the note. Later, Suzy, waiting at the restaurant, becomes disappointed.

Let *e* be the onset of Suzy's disappointment. Let *c* be Billy's trip. We have the right pattern of dependence: if Billy had not tripped, Suzy would not have become disappointed (because Billy would have made it to her house, seen the note, met her at the restaurant on time, etc.). According to any of a large number of counterfactual analyses *c* is therefore a cause of *e*. Now compare

No Note: Suzy and Billy want to have lunch together. But Suzy forgets to leave a note pinned to her front door, telling Billy where to meet her. On his way to her house, Billy trips over a tree root. He tries, but fails, to catch himself in time. He falls over. He bumps his head, and is knocked cold. Again, he never reaches Suzy's house. Later, Suzy, waiting at the restaurant, becomes disappointed.

No Note exhibits both a failure of dependence and a failure of causation: for if Billy had not tripped, he would have arrived at Suzy's house, seen no note, and gone back home in confusion; hence Suzy would have been disappointed all the same. It would therefore be absurd to count the trip a cause of Suzy's disappointment. But then the same verdict must hold for Disappointed, if the Intrinsicness thesis is correct. For the only intrinsic difference between the two situations concerns the disposition of the note—and this is a difference that on *any* sane view is extrinsic to the causal history of *e*, in Disappointed. In that situation, *e* has a no doubt complex causal history involving various psychological and perceptual episodes, as well as a number of other events that our story has left unspecified; if the counterfactual analyst is right, it also has among its causes Billy's trip. But even she will agree that it does not have among its causes the presence of Suzy's note on her front door. And so all should agree that whatever the causal history of *e* is, in Disappointed, that causal history is duplicated *exactly* in No Note. The Intrinsicness thesis then yields the conclusion that if *c* is a cause of *e* in Disappointed, so too is it a cause of *e* in No Note. The false consequent shows again how Intrinsicness conflicts with the thesis that counterfactual dependence suffices for causation.[17]

Why not blame Intrinsicness? After all, the idea that threat-cancelers such as Billy's trip are causes is not unattractive, at least to some (see Schaffer 2000b for a spirited and skillful defense of this view). Perhaps cases of threat-canceling simply reveal a curious way in which the causal structure of a process can fail to be intrinsic to it. (Lewis 2003 makes exactly this claim.) I think this position is defensible, but unattractive, in part because of the clear intuitive appeal of Intrinsicness, but more importantly because rejecting this thesis seems likely to render the problem of overdetermination intractable. At the very least, those who want to defend the claim that counterfactual dependence suffices for causation should recognize that they thereby render the problem of overdetermination vastly more difficult.

§7.3 Intrinsicness and causation by omission

We can obtain very similar results about causation by omission. Now, in order even to consider how to apply the Intrinsicness thesis to cases of causation by omission, we must take up an awkward question, which is whether to treat omissions as a kind of event, or rather to take reference to them as a mere manner of speaking, capable, perhaps, of being paraphrased away. In order to streamline discussion, I will simply take the first option (although it turns out not to matter: see my 2004 for discussion).

So we will assume that omissions are genuine events, just like kisses, collisions, and conversations. Well, not *just* like these other events, of course, for it is presumably of the essence

[17] It is perhaps worth mentioning that the provisional analysis of causation sketched above can be developed in a way that makes it clear that it does *not* conflict with Intrinsicness. See my 2004 for discussion.

of omissions and that they "happen" just in case no *ordinary* event of some appropriately specified type happens.[18] Suppose, for example, that in figure 10, **b** never fires:

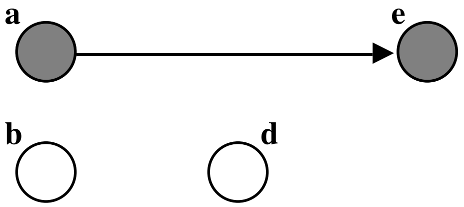

Figure 10

Then we might wish to credit its non-firing—the omission of its firing—as a cause of *e*, since if it *had* fired (within an appropriate time interval), **e** would not have fired (figure 11):

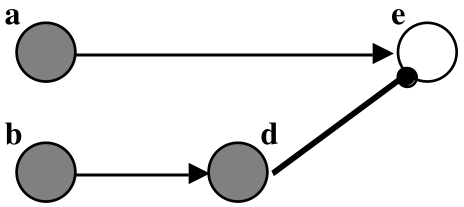

Figure 11

A problem now emerges that stands in the way of straightforwardly applying the Intrinsicness thesis to the events depicted in figure 10. Suppose we wish to collect together, into one event structure S, *e* together with all of its causes back to the time of **a**'s firing. Suppose further that we wish to include among these causes **b**'s failure to fire (along with other omissions, presumably). Our problem is to figure out exactly what contribution this omission makes to the intrinsic character of S. For an ordinary event this problem is easy to solve: We first

[18] Very occasionally, ordinary language obscures this point, as when I say, "Oops—I failed to leave the lights on", meaning of course not that some ordinary kind of event failed to happen, but rather that an ordinary event of the switch-flipping kind *did* happen.

locate the event, then figure out which particulars it involves, then figure out which properties they exemplify during the event's occurrence, singling out for attention those properties whose combined exemplification constitutes the given event. But how are we to execute this procedure, in the case of **b**'s failure to fire? Where, to begin with, should we locate this "event"? We might say that the omission is located where **b** is, during the relevant time interval. But the arbitrary nature of this answer becomes clear once we notice that **b**'s failure to fire can "occur" even if **b** does not exist. Suppose, for example, that shortly before time 0 something destroys **b**. We would still want to say that **b**'s failure to fire between time 0 and time 1 was a cause of *e* (that is, if we should ever countenance a case of causation by omission, we should countenance this one). But without **b** to serve as a peg on which to hang this omission, where are we to locate it?

Never mind. Suppose that we somehow solve this location problem. We still have not answered our original question: What does **b**'s failure to fire add to the intrinsic character of the event structure S? One defensible answer would be: nothing. For even if we agree to treat omissions as a species of genuine event, an excellent reason for sharply distinguishing them from kisses, collisions, and conversations is that, unlike these ordinary events, they lack intrinsic natures. If so, then the intrinsic character of S is exhausted by the intrinsic character of the ordinary events that make it up, together with the spatiotemporal relations among them.

Still, one could fairly defend a different answer. One could insist, for example, that once we have located **b**'s failure to fire as occurring throughout the spatiotemporal region R, we can specify the contribution that this omission makes to the intrinsic character of S as consisting in those intrinsic properties of R in virtue of which it contains no firing of **b** (together with the spatiotemporal relations region R bears to the other parts of S). This maneuver would guarantee that including **b**'s failure to fire as part of S makes a genuine difference to the intrinsic character of S, and so makes a difference to whether some alternative situation contains an event structure intrinsically matching S. Let us henceforth assume that this is how omissions contribute to the intrinsic character of event structures containing them.

Consider figure 10 again. It provides a canonical example of causation by omission. And now consider the situation depicted in figure 12, where neuron **d** is absent, and so not available to be stimulated by **b**, and so not available to send an inhibitory signal to **e**:

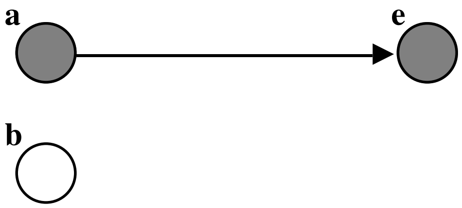

Figure 12

Here, **b**'s failure to fire is causally irrelevant to *e*. But on the proposal under consideration, the event structure S drawn from figure 10, and taken to *include* **b**'s failure to fire, will have in figure 12 a perfect intrinsic duplicate. So, if we try to extend the Intrinsicness thesis to cover the kind of causation by omission exhibited in figure 10, we get the incorrect result that in figure 12, **b**'s failure to fire is (still) a cause of *e*. More generally, we get the result that when the occurrence of an event of some type merely *could*—if circumstances had permitted—have been a threat to some outcome *e*, then the failure of an event of that type to occur causes *e*.

Notice the parallel with the discussion of threat-canceling: just as those who say that threat-cancelers are causes are forced, if they endorse the Intrinsicness thesis, to the absurd conclusion that events that "cancel" *non*-threats are nevertheless sometimes causes in virtue of doing so, so too are those who say that omitted threats are causes forced to say that omitted *non*-threats are sometimes causes as well. The fundamental reason is exactly the same in each case: The status of a threat *as such* often depends crucially on facts extrinsic to the causal history of the event being threatened.

I have stated these conclusions in an overly hard-nosed manner: If the Intrinsicness thesis is correct, then counterfactual dependence does not suffice for causation, and there is no causation by omission (i.e., situations that should provide examples of causation by omission, if *any* do, in fact fail to do so). I do not in fact believe these conclusions, so stated. That is because I think we should take seriously a third option, according to which there are different kinds of causal relation, one of which I call "production" and which is characterized (in part) by Intrinsicness, and another of which can be identified with counterfactual dependence.[19] Cases of overdetermination exhibit production without dependence; cases of threat-canceling and causation by omission dependence without production. I think a reductionist about causation should find this sort of causal pluralism entirely unremarkable. At any rate, if such a position is correct, then the key conclusion of this section needs a slight modification: counterfactual dependence may suffice for one kind of causation (which kind includes, I suppose, causation by omission), but if Intrinsicness is correct then it cannot suffice for all. In particular, it cannot suffice for the kind of causation exhibited by Suzy's throw in Suzy First. It is perhaps worth noting in passing that if this last claim is correct, then there is little interest in developing a counterfactual analysis of causation: for if dependence is the target then this is too easy, whereas if causation in all of its varieties is the target then it is impossible.

§8 Relevant similarity

Let me turn to the need for a refinement in the statement of the Intrinsicness thesis—a refinement hidden up to now in the claim that intrinsic match *in relevant respects* is what makes for similarity in causal characteristics.

§8.1 Need for the qualification

Why must Intrinsicness include this qualification? Because of the way we have used it to solve the problem of overdetermination. Recall the contrasting cases Suzy Alone, Billy Alone, and Suzy First. Suzy Alone and Billy Alone each contain a shattering-blueprint. We would like to say that Suzy's throw, in Suzy First, is a cause of the shattering because it belongs to an event structure that intrinsically matches the shattering-blueprint in Suzy Alone. But "matches" cannot

[19] More or less. There are complications, explored in detail in my 2004.

mean "perfectly duplicates", since Billy's rock causes gravitational effects that perturb the trajectory of Suzy's rock ever so slightly—and "ever so slightly" is quite enough to destroy perfect duplication. On the other hand, we cannot make due with mere intrinsic similarity, since *Billy's* throw, in Suzy First, belongs to an event structure that is intrinsically similar to the shattering-blueprint in Billy Alone, but cannot thereby qualify as a cause of the shattering. So we need it to turn out that this latter event structure is not intrinsically similar *in relevant respects* to the shattering-blueprint in Billy Alone, whereas the event structure containing Suzy's throw *is* intrinsically similar in relevant respects to the shattering-blueprint in Suzy Alone.

It would be disappointing if we could say nothing about what makes for similarity in relevant respects; for many, leaving this notion unexplained would give the Intrinsicness thesis a "whatever it takes" cast that would smack of triviality. That is not our situation, as I will shortly explain. But before doing so, observe that even if we were not able to—even if the explanation I offer fails disastrously—the charge of triviality would be inappropriate. For a weaker version of Intrinsicness—one that requires perfect intrinsic duplication rather than intrinsic similarity in relevant respects—is manifestly strong enough to secure the conclusions reached in the last section concerning threat-canceling and causation by omission. That's enough to show that Intrinsicness is non-trivial. (Of course, it's another question whether the weaker version would be worth defending, without the stronger: one reason to think not is that the weaker version is *too* weak to support a solution to the preemption problem.[20])

I will now present what I think is a promising and intuitive strategy for elucidating the notion of "relevant respects".

§8.2 Relevant similarity defined

Here is the key idea: Suppose we have some *e*-blueprint S, and some other event structure S' that is intrinsically similar to S. Such intrinsic similarity provides prima facie grounds for the claim that S' has the same causal structure as S; more exactly, for the claim that whenever *c* in S

[20] Thanks, here, to Jonathan Schaffer.

is a cause of *e*, then *c*' in S' is likewise a cause of *e*'. But these prima facie grounds can be overridden: in particular, if the intrinsic similarity between S and S' does not extend to those intrinsic characteristics of S *that matter to its having the causal structure it does*. It is exactly those characteristics that pick out the *relevant* respects of similarity. What remains is to provide some means for drawing them out. I will use our familiar cases to show how.

Begin by focusing on Suzy Alone. Suzy, unattended by Billy or anyone else, throws a rock at the window, breaking it. Imagine ringing very slight changes on this situation—not by adding anything to or subtracting anything from the environment, but by changing, in tiny ways, the intrinsic characteristics of the event structure S that consists of the shattering together with all of its causes back to the time of Suzy's throw. For example, we might change the exact angle of Suzy's throw, or the exact force that she imparts to the rock, or the mass of the rock, or the angular velocity it has during its flight, or its shape, or some of the details of its chemical composition, etc. Again, we might change the exact size of the window, or its thickness, or its internal chemical structure, or its color, etc. Or we might change several of these details at once. And now let us add two constraints: that in making these changes, we stay within the realm of nomological possibility, and that we leave causal structure invariant. That is, we require first that no change lead us to a situation that violates the fundamental laws at work in Suzy Alone. And we require second that the result S' of making slight intrinsic modifications to S have exactly the same causal structure as S.

The second requirement carries presuppositions that deserve to be brought out explicitly. I will take it that in order for S and S' to count as having the same causal structure, there must be some non-arbitrary way of imposing a one-one map from the parts of one to the parts of the other, such that *c* and *e* in S stand in some causal relation iff the corresponding events *c*' and *e*' in S' stand in that causal relation as well. What will make such a mapping non-arbitrary is for corresponding events to be reasonably intrinsically similar, and for them to occupy similar positions within their respective event structures. When such a non-arbitrary map exists, it need not be *unique*. Suppose for example that we modify our event structure S—the one drawn from

Suzy Alone—by having Suzy throw slightly more forcefully, thereby shortening the duration of her rock's flight. Then there will evidently be more than one way to non-arbitrarily match up the events constituting the slightly quicker flight of her rock in S' with the events constituting the slightly slower flight of her rock in S. No matter. What we need to require is simply that on any such way of non-arbitrarily imposing a one-one map from the events in S to the events in S', the two event structures will come out isomorphic with respect to their causal characteristics.

What we have done is to focus attention on all the nomologically possible ways of ringing slight changes on the intrinsic characteristics of S *that make no difference to its causal structure*. Quite obviously, there will be an indescribably large number of ways of doing so. Collect together their results into a set of nomologically possible alternatives to S; call this set the "blueprint-class" for S. There will be certain respects of similarity that all of the members of this blueprint-class share. In fact, these members will be intrinsically similar to one another, *except* in those respects that do not matter to their (shared) causal structure. So: they will be intrinsically similar to one another only in respects that *do* matter to their causal structure. So: for some event structure S' to be similar to S "in relevant respects" is just for it to be intrinsically similar to S in those respect in which S and other members of its blueprint-class are similar to one another.[21]

Let us apply this formula to our familiar cases. Start by letting S be the shattering-blueprint in Suzy Alone, and S' be the event structure in Suzy First consisting of Suzy's throw, the flight of her rock, the shattering, and the prior presence of the window: i.e., the event structure most intrinsically similar to S. Again, S' will not perfectly duplicate S, thanks to the perturbing presence of Billy's rock. But the slight intrinsic differences between S' and S will not be

[21] Certain constraints on the notion of a "respect of similarity" need to be in place, else this account of relevant similarity is a non-starter. Suppose the blueprint-class for S contains the event structures S_1, S_2, …. Then consider the highly disjunctive intrinsic property picked out by the predicate "perfectly duplicates S_1 or perfectly duplicates S_2 or …". If sharing of this intrinsic property counts as a genuine respect of similarity, then, obviously, the only event structures that will be similar to S in all those respects in which S and other members of its blueprint-class are similar to one another will be members of this blueprint-class. That won't do. So we evidently need to require that genuine respects of similarity correspond to not-too-disjunctive intrinsic properties. Happily, our ordinary notion of "respect of similarity" quite clearly meets this requirement.

differences in any respect in which S and other members of its blueprint-class are similar to one another. Hence S' will, by Intrinsicness, have the same causal structure as S; hence Suzy's throw in Suzy First counts as a cause of the shattering.

Now let S be the shattering-blueprint in Billy Alone, and S' be the event structure in Suzy First consisting of Billy's throw, the flight of his rock, the shattering, and the prior presence of the window: i.e., the event structure most intrinsically similar to S. As before, S' will not perfectly duplicate S. But this time, at least one of the intrinsic difference between S' and S *will* be a difference in a respect in which S and all other members of its blueprint-class are similar to one another: For it is true of every event structure in this blueprint-class that the rock strikes the window, whereas this is not true of S'. So S' does not count as intrinsically similar to S in relevant respects, and so Intrinsicness does not yield the conclusion that Billy's throw is a cause of the shattering.

§8.3 Causal efficacy

Let us pause for a brief digression. The ideas that led to an account of "relevant similarity" also give content to at least one notion of "causal efficacy". Let me point to the notion I have in mind by focusing on Suzy's throw in Suzy Alone. Intuitively, certain aspects of this throw matter to its status as a cause of the shattering, and certain other aspects do not: For example, it is important that the velocity her throw imparts to the rock is above some critical threshold, else the rock will not reach the window; likewise, it is important that the mass of the rock be above some threshold, else it will bounce harmlessly off the window. By contrast, the exact position of Suzy's pinky finger is not important, nor is the color of the rock. How shall we draw this distinction between *efficacious* and *inefficacious* aspects of Suzy's throw?

Here is an attractive way to do so: As before, let S be the shattering-blueprint in Suzy Alone, and construct its blueprint-class. Consider just those members of the blueprint-class that differ with respect to Suzy's throw, but *not* with respect to the intrinsic characteristics of the causes of the shattering that are contemporaneous with Suzy's throw. Equivalently: begin with S, and ring slight changes *only* on Suzy's throw (making necessary adjustments down the line to ensure nomological possibility); leave other factors such as the constitution of the window fixed. (Why hold these other factors fixed? Because if, for example, we weaken the window, then the momentum the rock needs to have in order to break it decreases; leading us to underestimate the causally relevant threshold value for this momentum.) The resulting subset of the blueprint-class for S picks out those intrinsic features of the throw that are efficacious: a feature is efficacious iff it is shared by the throws in each member of this subset.[22] For example, ringing changes on the *color* of Suzy's rock will result in a nomologically possible shattering-blueprint that is causally isomorphic to S; ringing (sufficient) changes on its *mass* will not.

It was important that we considered the pared-down situation provided by Suzy Alone, in carrying out this test for efficacy. For if we add stuff to the environment—even stuff that is, in fact, causally idle—the test fails. Consider:

Testy Demon: This case is just like Suzy Alone, save that a Demon is standing by, set on blocking Suzy's throw if (but only if) her rock is *red*. In fact, her rock is blue, and the Demon does nothing.

In Testy Demon, is the color of Suzy's rock—more exactly, its property of not being red—efficacious with respect to the shattering? Perhaps yes, in *some* sense; but there also seems a clear sense (the one I am trying to elucidate here) in which it is not: as before, it seems that all that matters about the throw are the mass of the rock, the angle and velocity imparted to it, etc. But the foregoing test does not yield this result. Change the color of the rock to red, while holding other contemporaneous factors fixed, and one does not arrive at a nomologically possible shattering-blueprint.

My tentative suggestion is that we treat causal efficacy in general as parasitic on causal efficacy, as exhibited in pared-down situations like Suzy Alone. Suppose we have some event *c*, and one of its effects *e*. To determine which aspects of *c* are efficacious with respect to its status

[22] As before, we should restrict attention to not-too-disjunctive intrinsic features. See note 16.

as a cause of *e*, we should proceed by (i) identifying some *e*-blueprint S that contains *c* as one of its initial constituents (i.e., *c* is part of S, and no event earlier than *c* is part of S); (ii) identifying some (possibly distinct) event structure S' that intrinsically matches S in relevant respects, and that exists in "pared down" circumstances containing no back-up processes whose initiation is sensitive to the manner of occurrence of *c*'; (iii) identifying the efficacious aspects of *c*' according to the foregoing test. We can then take the efficacious aspects of *c* to be exactly the same as the efficacious aspects of *c*'.[23]

§9 A potential counterexample, and some lessons

Trouble for the Intrinsicness thesis comes in the form of a kind of example that appears to draw forth firm intuitions clearly at odds with that thesis.[24] This section presents the example (due to Steve Yablo), along with a diagnosis that goes some way towards preserving Intrinsicness in the face of it; this diagnosis will suggest some deeper lessons about the way in which Intrinsicness functions to structure at least one of our central causal concepts.

§9.1 The example

Here is the case, which we imagine to take place in a Magic world, in which wizards regularly cast spells (we'll see shortly that it's a mere convenience to rely on "magic"):

[23] It may have been noticed that my account of the intrinsic character of causation makes no provision for perfectly coincident events that (i) are numerically distinct, because they differ in which of their categorical properties are *essential* to them; and that (ii) have different causes and effects. (See for example Lewis 1986c and Yablo 1992.) For example, when I shut a door by slamming it, some would have us distinguish two closings of the door: one which is *essentially* a slamming, and one which is only *accidentally* a slamming. Those who draw such distinctions will typically want to go on to distinguish the causes and effects of such perfectly coincident events. As stated, Intrinsicness rules out such causal distinctions, since any event structure containing one of a pair of perfectly coincident events will have a perfect duplicate which contains the other. One could, perhaps, modify Intrinsicness to allow for causal distinctions between perfectly coincident events, but I think a much better course is to argue that any philosophical work to be done by such distinctions can be better carried out by a careful analysis of causal efficacy, perhaps along the lines sketched in this subsection. I explore this issue in much more detail in my 2004.

[24] I bypass here discussion of other alleged counterexamples, e.g. the "trumping" cases considered by Schaffer in his 2000a, and the "switching" cases examined in my 2000. By contrast with the kind of case considered in the text, none of these cases strike me as particularly rich in probative value. But for full discussion of these and other examples, see Hall and Paul 2002, and my 2004.

Meddling Wizards: Suzy throws a rock at a bottle. Two wizards, Merlin and Morgan, are standing by. Morgan casts a spell to make her rock disappear in mid-flight. Merlin casts a spell to make a rock appear, flying towards the bottle. It so happens that the two spells dovetail so perfectly in their effects that the instant Suzy's rock disappears, its place is taken by Merlin's rock. What's more, Merlin's rock is intrinsically just like Suzy's, and flies towards the bottle in just the way Suzy's would have, had it not disappeared. The bottle shatters.

It seems perfectly clear that in Meddling Wizards, Suzy's throw is *not* a cause of the shattering. But now consider a variant, in which the two wizards are absent, Suzy throws in exactly the same way, and the bottle shatters. That variant contains a shattering-blueprint S that is, we are supposing, duplicated *exactly*, down to the minutest detail, in Meddling Wizards (for that is how perfectly the two spells dovetail). By Intrinsicness, then, Suzy's throw in Meddling Wizards should be a cause of the shattering.[25]

If we are careful about how to describe the laws in this world, we will see that we need to consider a quite different causal gloss that could be put on the example—one that is in fact consistent with the verdict that Suzy's throw causes the shattering. For it is possible that the causally loaded locutions with which we described Morgan's and Merlin's spells illegitimately prompted the troublesome intuitions about the case. So suppose that, as is appropriate, we describe the spells in non-causal terms, by simply specifying what events must (given the laws) take place, consequent upon their casting. Presumably, the right such "non-loaded" description is the following: (i) In a situation in which neither spell is cast, the rock follows a certain trajectory T; (ii) in a situation where Morgan's but not Merlin's spell is cast, the rock vanishes at a certain point in its trajectory; (iii) in a situation where Merlin's spell is cast, the rock follows trajectory T (regardless of whether Morgan's spell is cast). We could put (iii) more cagily, thus: in a situation

[25] One might worry that the shattering-blueprint is *not* duplicated exactly, since Merlin's rock is not (numerically) the same as Suzy's. That's a red herring: just rewrite the example so that it doesn't involve a persisting object like the rock, but rather some other sort of process. For example, let Suzy light a fuse, Morgan cast a spell to snuff out the traveling spark, and Merlin cast a spell to reignite it.

where Merlin's spell is cast, a certain trajectory-shaped region of spacetime is rock-filled. It is now not so perfectly clear what the causal structure of Meddling Wizards is. In particular, it is not clear why we should reject the following gloss: the effect of Merlin's spell is merely to *neutralize* Morgan's spell. And that is a gloss which preserves Suzy's throw as a cause of the shattering.

In the end, I do not think this gloss is appropriate; I think rather that the original verdict stands. But it is an exceedingly useful question to ask *why* it is inappropriate. Do not say: because it is sheer coincidence that Morgan's and Merlin's spells dovetail in the way that they do. For that is a feature of the case easily erased: let God have ordained, since the dawn of time, that the two wizards would cast their spells just so. Do not say: because the shattering counterfactually depends on Merlin's spell, and not on Suzy's throw. For again, that is a feature easily erased, by addition of suitable extrinsic details (e.g., in the form of an even more powerful wizard who will act so as to guarantee that the shattering occurs in just the manner it does, should Suzy, Morgan, and Merlin collectively fail to do so). And do not say: because if Suzy had thrown, say, a watermelon, then it would still have been a rock that shattered the window (so that there is no counterfactual covariation of the manner of the shattering with the manner of her throw). For it's easy to *introduce* such covariation in a way that does not alter our intuitions about the case (e.g., let the more powerful wizard stand ready to intercede, if Merlin fails to produce an object perfectly matching whatever Suzy threw). But do not, finally, dismiss the case as having nothing to teach us. On the contrary: there is a natural explanation for the strength—even in the face of the alternative gloss—of the intuitive judgment that Suzy's throw is not a cause of the shattering.

To see what this natural explanation is, it will help to represent Meddling Wizards abstractly, using our handy neuron diagrams:

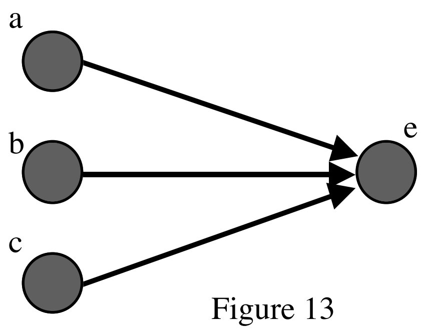

Figure 13

Neurons **a** (Morgan), **b** (Suzy), and **c** (Merlin) fire simultaneously, at time 0; their signals reach **e** at time 1; an instant later, **e** fires. Let the laws dictate that (i) **e** will always fire upon receiving a signal from **c**; (ii) if **c** does not fire, then **e** will fire upon receiving a signal from **b**, provided it does not also receive, simultaneously, a signal from **a**; (iii) **e** will not fire unless it receives a signal from at least one of **b** and **c**.

It appears that we could describe the events depicted in figure 13 in either of two ways: On the one hand, we could say that *c* alone causes *e*, while *b* is not also a (symmetrically overdetermining) cause because of the inhibiting signal from **a**. On the other hand, we could say that the signal from **c** helps bring about *e* only by neutralizing the inhibitory signal from **a**, thus allowing the signal from **b** to do its work. What considerations can we adduce that would favor one description over the other?

§9.2 The intrinsic character of prevention

The crucial consideration is brought out by figure 14:

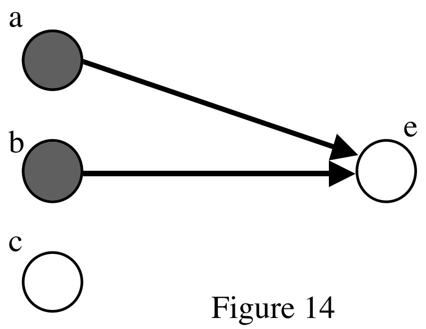

Figure 14

As before, **a** and **b** fire at time 0, sending signals that reach **e** at time 1; but because no signal from **c** reaches **e**, **e** fails to fire an instant later. So the signal from **a** prevents **e** from firing. But there is a second, distinct kind of preventative relationship that these events manifest as well: the signal from **a** prevents the signal from **b** from causing **e** to fire. I claim that this second relationship is intrinsic to the event structure that manifests it. That is, consider the event structure S consisting of (i) the firing of **a**, together with the passage of its signal to **e**; (ii) the firing of **b**, together with the passage of *its* signal to **e**; (iii) the presence of neuron **e** from time 0 to time 1. It is *not* a fact intrinsic to S that **e** fails to fire an instant after time 1; hence it is not a fact intrinsic to S that the signal from **a** prevents **e** from firing (since **a** prevents **e** from firing only if **e** does not fire). But the claim that the signal from **a** prevents the signal from **b** from causing **e** to fire does not entail the claim that **e** fails to fire an instant after time 1. So it is tenable to suppose that the fact that the signal from **a** prevents the signal from **b** from causing **e** to fire *is* intrinsic to S. I propose—with a slight qualification shortly to come—that it is not merely tenable, but true.

A full defense and elaboration of this proposal belongs elsewhere. Suppose this work done, so that we have available to us an upgraded version of Intrinsicness that applies not just to causation but to prevention (in the way indicated in the last paragraph). Let us see how to apply it to the analysis of figure 13. Observe that figure 13 depicts, among other things, a perfect intrinsic duplicate of S. In addition, it contains a perfect duplicate of the *e*-blueprint exhibited in figure 15:

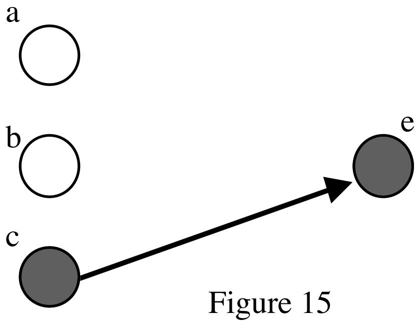

Figure 15

Making use of the <u>Intrinsicness</u> thesis, we should therefore conclude that in figure 13, (i) the signal from **a** prevents the signal from **b** from causing **e** to fire (thanks to the match between figures 13 and 14); and (ii) the signal from **c** causes **e** to fire (thanks to the match between figures 13 and 15). But given (i), it cannot *also* be that the signal from **b** is a cause of *e*. So *c* is the *sole* cause of *e*.

A problem remains. What we have done is to take note of certain causal structures in figures 14 and 15, and to make use of <u>Intrinsicness</u> to project these causal structures into figure 13. But why is it not just as legitimate to take note of the causal structure in figure 16, and use <u>Intrinsicness</u> to project *it* into figure 13?

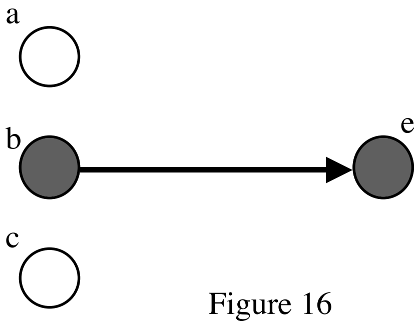

Figure 16

We know that we cannot project both, on pain of inconsistency. But what favors projecting the instance of prevention, as opposed to the instance of causation?

A proper answer needs to weigh the costs of each alternative. On the one hand, if we project the causal structures in figures 14 and 15, then we must say, *contra* <u>Intrinsicness</u>, that even though figure 13 contains a perfect duplicate of the *e*-blueprint found in figure 16, *b* is not a cause of *e*. On the other hand, if we insist that *b is* a cause of *e*, then we must say—also *contra* <u>Intrinsicness</u> (as modified to accommodate prevention)—that even though figure 13 contains a perfect duplicate of the prevention-blueprint in figure 14, *a* does not prevent *b* from causing *e*. So far the score is even. But there is an *extra* cost to this latter alternative. For taking it *also* requires us to assign to *c*, in figure 13, a novel causal role: namely, that of "neutralizing" the inhibitory signal from **a**. That doing so is inappropriately *ad hoc* emerges when we observe that there are no *other* circumstances, qualitatively distinct from those exhibited in figure 13, that would attest to such a role. Thus the balance tips in favor of the first alternative.

§9.3 Intrinsicness and the function of our causal concepts

The picture we arrive at is this: Causation—or at least, one important kind of causal relation—is governed by the Intrinsicness thesis. But it is governed loosely, in that conflicts can arise *from that thesis itself*: very occasionally, Intrinsicness will offer contradictory advice about what causal structure to assign to some sequence of events. Resolving such conflicts requires appealing to higher principles, principles that will pick out the particular bit of advice that should be heeded. The argument of the last paragraph tacitly appealed to such principles—without, to be sure, articulating them at all explicitly. What, then, are these principles, and what grounds them?

I must leave these as open questions. But not without suggesting what seems to me an attractive line of investigation. Consider that reductionists like me—philosophers, that is, who hold that the *fundamental* causal structure of the world is given by what happens (what the complete physical state is, at each moment of time), together with the fundamental laws that govern what happens—face an excellent question: Why, by our lights, should there be any *need* for a concept of a causal relation that applies to anything less grand than a complete physical state? A promising answer is that given our relatively impoverished epistemic circumstances, such a concept can play the crucial role of allowing us to capture—in a manner that is both approximate and piecemeal but, above all, *manageable*—facts about the fundamental causal structure of the world.

Think of the matter this way: As we amass more and more facts of the form "*c* caused *e*", "*a* prevented *e*", "*a* prevented *c* from causing *e*", etc., we build up, bit by bit, a better and better picture not just of what happens in our world but of what the fundamental laws are that govern what happens. I suggest that in order for our causal concept to best play this role, it should be governed by something like the Intrinsicness thesis: For then we will be able to use it to delimit the causal features of novel situations, by discerning in them already-familiar patterns in the form of event structures intrinsically similar, in relevant respects, to ones whose causal characteristics we have antecedently settled. Put another way, we want a concept of causation that will allow us to describe the nomological shape of the world as efficiently and clearly as possible in terms of *repeatable causal structures*. Given that aim, it makes exceedingly good sense to employ a concept governed by Intrinsicness. For without it, it seems that whenever we encountered some slightly novel situation, differing in some small respect from those with which we are familiar, we would not be entitled to use our experience of similar situations to make even a provisional guess as to its causal structure. Rather, we should have to start afresh. Imagine, in this regard, being familiar with innumerable cases like Suzy Alone, and then encountering Suzy First: Is it really plausible that we would start afresh *here*, simply *ignoring* the deep intrinsic similarities between this new case and the old ones? If so, then surely—given its *counterfactual* structure—our first guess as to the *causal* structure of Suzy First should be that it is a case of symmetric overdetermination. Nobody, but *nobody,* makes *that* mistake.

When, finally, we face one of those rare occasions where this thesis endorses contradictory descriptions, we should settle on that one which best fits our overarching aims. In the case of figure 13, we violate these aims if we assimilate its causal structure to that of figure 16. For in doing so we are forced to introduce a causal structure (the "neutralization" of the **a**-signal by *c*) that is not only novel, but *unrepeatable*, outside of circumstances exactly like those of figure 13. Assimilating the structure of figure 13 to that of figures 14 and 15, by contrast, carries no such cost.

§10 REFERENCES

Collins, John, N. Hall, and L. A. Paul eds. (forthcoming 2003). Causation and Counterfactuals, Cambridge: M.I.T. Press.

Dowe, Phil (1992). "Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory", Philosophy of Science 59: 195-216.

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

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

Hall, Ned (2000). "Causation and the Price of Transitivity", Journal of Philosophy 97, pp. 198-222.

Hall, Ned (forthcoming 2003). "Two Concepts of Causation", in J. Collins et. al. eds. (2003).

Hall, Ned (forthcoming 2004). Causation, Oxford: Oxford University Press.

Hall, Ned and L. A. Paul 2002. Causation and the Counterexamples: A Traveler's Guide. Ms.

Lewis, David (1986a). Philosophical Papers, Vol. II, Oxford: Oxford University Press.

Lewis, David (1986b). "Causation", with "Postscripts", in Lewis (1986a): 159-213. Originally published in Journal of Philosophy 70 (1973): 556-67.

Lewis, David (1986c). "Events", in Lewis (1986a), pp. 241-269.

Lewis, David (forthcoming 2003). "Causation as Influence", in J. Collins et. al. eds. (2003).

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

Maudlin, Tim (1996). "A Modest Proposal Concerning Laws, Counterfactuals, and Explanation", unpublished ms.

Menzies, Peter (1999). "Intrinsic versus Extrinsic Conceptions of Causation", in Howard Sankey ed. (1999).

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

Sankey, Howard ed. (1999). Causation and Laws of Nature, Kluwer.

Schaffer, Jonathan (2000a). "Trumping Preemption", Journal of Philosophy 97, pp. 165-181.

Schaffer, Jonathan (2000b). "Causation By Disconnection", Philosophy of Science 67, pp. 285-300.

Tooley, Michael (1990). "Causation: Reductionism versus Realism", Philosophy and Phenomenological Research 50, Supplement: 215-36.

Yablo, S. (1992). "Cause and Essence", Synthese 93: 403-449.