IX*—WOULD CAUSE

by Adam Morton

I am going to discuss the interaction between the subjunctive or (so called) counterfactual conditional and various causal ideas and idioms. My larger intention is to look at the place of the concept of causation among other modal ideas. I assume that there is a more or less unitary concept of causation, which we express by saying 'causes', 'because of', 'made happen', and by using verbs whose presuppositions are causal (e.g. kill, break, make). This is not to say that what is involved in the use of such words need ever consist just in causality. And I assume that causation is itself a very particular causal concept; that is, that it is just one of a family of ideas, others being expressed by the subjunctive conditional and the idea of a law of nature, which all together give us our picture of the world as causally ordered and potentially intelligible. I take it that efforts on a number of fronts to unravel the ideas of a law of nature and of a subjunctive conditional are going as well as one could expect. And that if we had a *good* understanding of these we would have *most* of what Hume wanted in order to understand our 'reasonings concerning matter of fact', without dealing with the idea that 'causes', as contrasted with these other devices, expresses. What is left out when we leave out causation?

In the more unproblematic parts of what follows I take for granted standard possible worlds analyses of the counterfactual, and I allow myself to tinker with them in only a superficial way, in order to have something of some solidity to measure cause up against. I shall look at sentences of the form 'if p then e_1 would cause e_2', and some related forms, in order to see what the 'would cause' adds to the 'if'. There is an obvious puzzle about the role of causation here. If one says 'if you step on it, it'll break' one has already described its breaking as a causal consequence, a result, of stepping on it. What is added by saying 'if you step on it, that'll *cause* it to break'? Something surely is. And something subtler than what 'stepping on it caused it to break' adds to 'you stepped on it and it broke.'

* Meeting of the Aristotelian Society held at 5/7 Tavistock Place, W.C.1 on Monday, March 16, at 6.30 p.m.

I believe that following out this question throws some light on the conditional. But before I get down to work on it, some rather vague remarks on causation and conditionals may help explain why I take things in the direction I do.

David Lewis is responsible for an elegant treatment of causation. Lewis' idea was to start with the simplest of Hume's definitions of cause, that 'e_1 causes e_2' is true when 'if e_1 had not occurred e_2 would not have occurred', to take the 'if' seriously as a counterfactual conditional, and then to see what refinements are necessary to make it fit our intuitions about the causal relation. Suppose that this were an adequate account of 'causes'. Then 'if p then e_1 would cause e_2' would take the form 'if p then if e_1 had not occurred e_2 would not have occurred'. This is rather cumbersome, and one would naturally look at the simpler 'if p then if e_2 had occurred e_1 would have occurred', even though it is not equivalent when the 'if' is counterfactual, and also at 'if p then if e_1 had occurred e_2 would have occurred'. Without the 'if p' these won't have any force as even the beginnings of analyses of 'causes', since if e_1 and e_2 do actually occur 'if e_1 then e_2' and 'if e_2 then e_1' are hard to make sense of in many cases. But this is a fact about 'causes' that won't affect 'might cause'. For 'might cause' it is perfectly reasonable to ask whether 'if there had been oxygen present that spark would have caused an explosion' is most closely related to 'if there had been oxygen present then if that spark had occurred an explosion would have followed', or 'if there had been oxygen present then an explosion would only have occurred if that spark had occurred', or 'if there had been oxygen present then if that spark had not occurred no explosion would have followed'.

Let this idea remain undeveloped for a while, and consider another. Usually in a counterfactual conditional the antecedent represents something happening earlier in time than the consequent does. If she had married me she'd be miserable today. But note that there are a number of cases in which although both antecedent and consequent intuitively represent events occurring at definite times, the event represented by the antecedent is represented as occurring *later* than that represented by the consequent. For example, 'if she arrives by 1.30 she will have —by some lucky chance—managed to catch a taxi on the street' or 'if she arrives by 1.30 she will have had to catch a taxi

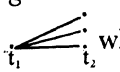
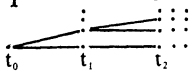
on the street', or 'if she arrives by 1.30 it will be because she (managed to) catch a taxi on the street'. Usually some little interjection is needed 'by some lucky chance she managed', 'will have had to' (read with a special intonation), 'it will be because', to indicate that there is something unusual about the conditional. The last of these seems to me particularly significant. A causal idiom is introduced partly in order to make explicit that the conditionality involved is one that can connect later events to earlier ones.

These two 'ideas' are meant just to hint at some larger significance behind the tinkering about that now follows, and to soften you up for the speculative conclusions that follow that tinkering. The actual argument depends on an analysis of what is responsible for three problems of the now-classical Stalnaker-Lewis analysis of the counterfactual.

Time Problems (the 'then' in 'if-then'). In a subjunctive conditional the time of the antecedent is, as I remarked, usually earlier than that of the consequent. One can explain why this should be, if one accepts the general spirit of the Stalnaker-Lewis account and combines it with a branching-worlds model of the proximity of one world to another. Let me describe each of these, only to the extent that it is necessary for my purpose.

The central idea of the Stalnaker-Lewis account must by now be very familiar. It is that a conditional 'if p were the case then q would be the case', is true at a world *w* if 'q' is true at all worlds which are (a) such that p is true at them and (b) no more remote from *w* than the nearest world at which p is true. Or, simplified, the conditional is true if the consequent is made true by all variations on actuality which are just large enough to make the antecedent true. The talk of arbitrary worlds *w* and the peculiar idiom 'true at w' is there to allow us to handle iterations of conditionals, such as 'if p then if q then r', in the way that Kripke taught us. The main virtue of the account is that it explains why 'if I drop this egg on the floor it will break' is true in spite of the fact that there are many (physically possible) worlds in which I drop it and it does not break (e.g. one in which the floor has been covered in cushions): such worlds are more remote than ones in which the situation is sufficiently like the actual facts that the egg breaks when dropped.

Which conditionals are then as a matter of fact true will depend on which worlds are as a matter of fact more and less remote from which others. It is natural to try to make sense of this in terms of an image of the world evolving in time, at each moment sprouting a set of alternatives, representing the ways things might have gone at that point.

At any point in time various things can happen next, some more likely than others. So we have  where the points at t_2 displaced upwards more greatly from t_1 represent the less likely evolutions from the state of the world at t_1. (For simplicity, I consider time to be a discrete series of moments.) And after each of these possible moments other possibilities open up, of various likelinesses, so that from any given present there is a whole ramifying tree of future developments  and so on. A world is a complete branch of the tree, from beginning to end. Events are propositions whose truth is determined at one moment in time; at any rate that is how I shall use 'event', so that I may speak of events being true at worlds and times. There is no assumption that the actual world is always the most likely branching.

Let us assume that the basic determinant of the separation of two worlds is the 'size' of their accumulated divergences from the first point in time at which they differ. That is, take for granted the idea of one branching between two adjacent points of time being more unlikely, more of a divergence, from the state of things at the earlier time, than another, and then measure the separation of two worlds by the magnitude of these accumulated step-by-step divergences. (The divergence of a world w up to a point t_n from an initial difference at t_0 can thus be thought of numerically, as $\sum_{i=0}^{n} d(i)$, where $d(s)$ is the divergence, the 'vertical' displacement, at t_s of the branchings representing w. Note that this 'measures' the divergence of w from the most likely 'horizontal' continuation at t_0. One might want a subtler measure of the divergence of w from some unlikely continuation, as the actual world might be. The numerical measures may seem pointless in the midst of all this hand-waving, but they do bring out both the fact that I shall exploit later, that there are different Stalnaker-

esque definitions of the separation of worlds that make equally good sense on this model, and that the divergence between two worlds *in toto* may be incalculable, the $\sum_{i=0}^{\infty} d(i)$ may be infinite, although the divergence up to any point in time will be a calculable figure. Throwing away the pretence of quantification, it still seems very likely that one could say of two worlds up to some (or even any) future time which one is more removed from actuality, but find it impossible to tell, there just may be not intelligible fact to the matter, which considered in its total extension in time is more different from the actual course of things.)

Now, the point of all this. Suppose that we have a counterfactual conditional 'if e_1 had occurred at t_1, then e_2 would have occurred at t_2', where t_1 is earlier than t_2. Then we want to look at departures from actuality before t_1 (not too long before t_1, on the usual meaning of the conditional) which lead to e_1 at t_1; we want to find the smallest divergences from actuality up to t_1 (or, hearing 'if' slightly differently, up to t_2) of such worlds; and we want to check that in all such worlds with no greater than this minimal divergence from actuality, e_2 is true at t_2. This isn't hard to do. We just follow all paths branching from the branch representing the actual world shortly before t_1 and do some calculation. Next consider a counterfactual 'if e_2 had occurred at t_2, then e_1 would have occurred at t_1', where again t_1 is earlier than t_2. To see if it is true we have to look at all worlds which lead to e_2 at t_2 and then work backwards to their departures from actuality, and then work forwards again, adding up the size of their divergences and checking which ones make e_1 true at t_1, but presumably continuing on past t_1 to t_2 in adding up the divergences. One has to go all the way back and then all the way forward again; actually, in order to identify the possible chain of events one is tracking one probably has to have a *three*-way search, starting from a prior time and going forward on rough bundles of chains till one finds e_2 at t_2, then working backwards on particular chains to find their points of departure, then working forward to look for e_1 at t_1 and continuing on to t_2 summing up divergences. No wonder the mind sometimes boggles when asked to consider a conditional with the times backwards. One hasn't the simplifying option of cutting off one's summation at t_1, and one has the complication of having to grope back a longer way

to find the point of departure. (And one cannot make the tacit assumption, which many conditionals encourage, that only worlds departing from actuality not long before the time of the antecedent need be considered. For by the nature of the case one is committed to looking at worlds departing at least as far before the antecedent as the time of the consequent.)

I am glossing over a lot. Of course divergences cannot be literally added up like numbers; of course time does not proceed step by step; of course events, let alone propositions in general, don't have absolutely exact time indices. But we do think out the acceptability of conditionals in something like this way; we proceed along possible courses of events toting up how likely they are, development by development. Subtly different ways of doing this represent subtly different conditionals, presenting themselves as 'if', which are indicated by details of context and wording. And in terms of this it is clear that while 'backwards' conditionals, in which the antecedent refers to something as happening later than the possibility the consequent refers to, *can* be understood, it is also clear that they are much harder to understand, that one has to put more work into thinking out whether they are true or false. The next question is: when one wants to assert a relation of conditional dependance of a later possibility on an earlier one, what linguistic devices might make it easier to understand what is said?

The obvious way to make backwards conditionals easier is to make them less backwards. The source of the awkwardness of the backwards conditional is the number of chains of events that can lead to an event e_2 being true at some later time than a vaguely-specified earlier time; there are just too many ways e_2 can come about. The way to prevent this is to restrict the possibilities, to describe fairly restrictive conditions centring on some earlier point in time, and then to say: if these conditions held at t_0 then if e_2 occurred at t_2 it would be proceeded by e_1 at t_1. With the right choice of conditions there will be few enough worlds leading from their satisfaction to the truth of e_2 that it is easy to check whether they all make e_1 true. (Note well: the conditions have to be restrictive of how e_2 can occur, the earlier e_1 is then squeezed between the conditions and e_2.)

An example. It is hard to understand 'if that boulder will roll over the cliff (in two minutes) you will push it (now)', even if

rephrased more idiomatically as 'if that boulder rolls over the cliff, you will have pushed it'—hard to understand in the sense that given a lot of information about the situation one would still be unsure whether it was true. Now if we insert a third 'backing up' condition, we get conditionals such as 'if that boulder is poised delicately enough, and there's no wind blowing and no-one else of a mind to push it, then if it rolls over the cliff you will have pushed it'. Compare also 'if there's oxygen present and it's not raining and there are no electrical discharges, then if the match lights you will have struck it'. (Note that some of the prior conditions are meant to ensure that there are some histories that lead to both antecedent and consequent, and others are meant to ensure that all histories leading to the antecedent (at the later time) lead through the consequent (at the earlier time).) These are clearly more easily understood; in fact they are perfectly natural.

One would usually say something simpler than these backwards conditionals with their careful limiting conditions. One would say 'if that boulder is poised delicately enough (etc) then if it rolls over the cliff it will be because you pushed it', or 'if that boulder is poised delicately enough (etc), a push from you will send it over the cliff', or 'if there's oxygen, and it's dry and there are no sparks, then the way to light the match is to strike it'. Or others; they are all different, in various subtle ways, and all add something to the 'if C at t_0, then if e_2 at t_2, e_1 at t_1' pattern. They are all causal idioms, though, and all show a form of Mackie's famous INUS conditions. They all say of an 'effect' event that given a 'cause' event and some conditions, the effect will only occur if the cause is added to the conditions. The comparison with what Mackie says is not perfect, in part because he is dealing with 'causes' rather than 'would cause', and I shall not press it. But it does strengthen my conviction that what has been added to the conditional in these formulations has something essential to do with cause. To put it briefly, an essential part of the content of 'it would be because' is that it makes a conditional, whose time references are backwards, more manageable, by imposing conditions that limit the stretch of time that has to be traced out (backwards and then forwards) to a manageably short interval and a manageably small variety of possible histories.

Two further observations before I move on to another compli-

cation of the conditional. First, one often leaves out the limiting conditions. One says simply, 'if the match lights it will be because you struck it', 'if the boulder rolls it will be as a result of your pushing it'. And evidently very often when these do seem appropriate, it isn't the case that striking the match or pushing the boulder are really the *only* ways of getting them to light or roll. There is an *implicit* antecedent condition here, pretty clearly, to the effect that there are conditions existing soon before the striking or the pushing might occur, which serve to restrict the possible chains of events suitably. And this is surely also essential to the function of the 'because' in these sentences: to assert the existence of such counterfactual INUS-like conditions. (So with enough work you can turn Mackie into Lewis.)

The other observation is that we now have some hold not only on 'it would have been because' but also on 'would cause'. If e_1 would cause e_2 then if e_1 occurred e_2 would, and e_2 would occur because of e_1. This looks rather like saying that e_2 would occur if and only if e_1 (where the if was temporarily frontwards in one direction and backwards in the other), but to say this would be to ignore the role of the antecedent conditions, and would also be misleading because the two 'if's do not have quite the same force, as the next section should show.

Only if The counterfactual conditional lacks many of the logical properties of the material conditional. That is, important inference patterns, such as contraposition and transitivity, can seem dubious when the 'if' is subjunctive, and formal counter-examples to them can be constructed if one does one's model theory in Stalnaker's way. The reason usually comes down to the fact that the consequent of the conditional need be true in only those worlds in which the antecedent is true and which are least different from actuality. (I must stress again that it is this feature that is responsible for the most appealing consequences of the theory.) For example, contraposition fails because 'if not p then not q' is true and 'if q then p' is false if in the worlds nearest to actuality not p and not q are true, then in worlds somewhat more distant not p and q are true, and only in worlds yet more distant are p and q both true. Then since in the nearest worlds in which not p is true (but not in *all* worlds in which not p is true) not q is true, 'if not p then not q' is true, but in the nearest worlds

in which q is true not p rather than p is true, so 'if q then p' is false.

This can be plausible with the real English if. Suppose that you're lazy, averse to pushing the boulder before us, which is in any case pretty solidly stuck to the edge of the cliff. It is however *just* possible that that man with the tractor will drive into the boulder, sending it rolling over the edge; it is more likely, at any rate, than your having the energy to push. However if you did push, it would roll. So it is true—isn't it?—that if you don't push it won't roll, but false that if it rolls you've pushed it. (Because if it rolls it'll be because of the tractor.) (If you think that in this case it is false that if you don't push it won't roll, you can modify the case by picking something less likely than the tractor, to play that role, and so on *mutatis mutandis*, and make a counterfactual that does work for you.)

So far so good, but one seems also to have counterexamples to the inference from 'if q then p' to 'if not p then not q', when the nearest worlds are such that p and q, the next nearest such that not p and q, and only then come worlds in which not p and not q. Then 'if q then p' is true, since the nearest q-worlds make p true, but 'if not q then not p' is false because the nearest not-p-worlds also make p true. For example: you're pretty eager to push and the boulder is poised very precariously on the edge, but your enthusiasm is not a reliable thing and you might be daunted by the idea of pushing before you actually began, but then there's quite a wind blowing so almost certainly (not completely certainly) the boulder will roll anyway. Then it seems clear that 'if you don't push it won't roll' is false, and pretty clear that 'if it rolls, you pushed' is true, since you were very likely to push and it was very likely to roll and it was even very likely that the push would help the rolling. There *is* something a little disturbing about the conditional, no doubt connected with the lack of causal connection between antecedent and consequent. But more worrying is the observation that the backwards conditional 'if it rolls you pushed' ought to be rephraseable as 'it rolls only if you pushed it'. And this seems quite obviously false. As does 'if it rolls, it'll be because you pushed'. The shade of causal connectedness implicit in the simple 'if' is more central with these variants. They seem to bring out a 'causal' sense of 'if' which imposes stronger conditions.

And one part of those stronger conditions is surely expressed by a natural reading of the 'only' in 'only if'. The *only* way the boulder can roll is by your pushing; there's no other way. But in the example pushing is not the only route to rolling, though it is the most accessible one. But saying this goes against Stalnaker's basic idea, that one must not insist that the consequent be true under *all* conditions that would satisfy the antecedent.

This certainly shows how hard it is to work out patterns of inference in English. The senses don't stay still long enough to get the equipment focussed. A subjunctive conditional often will be most naturally read as saying what Stalnaker or Lewis would have it say, and then a little shift of tenses, a contraposition, an 'only', or a 'because', will shift the shading of the 'if' slightly, so that in the conclusion of an inference it has a slightly different sense than in the premises. What interests me now, of course, is that the shift of meaning I've been discussing adds a causal colouring to the 'if', it requires a causal connection between antecedent and consequent and is naturally expressed by 'it would have been because of'. The importance of this is increased by the coincidence in cases like this of a shift of meaning signalled by 'because' with one signalled by 'only'. For 'only' reverses the order of antecedent and consequent, typically resulting in a 'backwards' conditional, at the same time as increasing the force of the conditionality, inserting more causal necessitation into it.

But this coincidence is just what one would expect from the last section. A backwards conditional will be most manageable when there is an implicit prior condition limiting the range of possibilities and the stretch of time over which they can develop; given this restriction, there is no need also for a Stalnakerian limitation to nearest possible worlds *within* the scope of the restriction—one can look at all chains of events which might lead from the initial limiting conditions through the 'cause' to the 'effect'. Or to put it differently: both 'only if' and 'it would have been because' allude to an underlying triadic conditional 'if C (at t_0) then if e_2 (at t_2), e_1 (at t_1)' where the first if is taken in Stalnaker's way and the second is unrestricted, it requires that all possible routes (among those selected by the 'least departure' principle applied to C) which lead to e_2 do so by passing through e_1. I think that this double conditionality, where the force of the two if's is different, is the essence of the causal.

If p then if q then r We now have something like a definition of 'would cause'. e_1 would cause e_2 under conditions C if (a) if C obtained then if e_1 occurred e_2 would occur, where both 'if's are Stalnakerian (I'm not really sure they are), and (b) if C obtained then if e_2 followed it would be because of e_1, where this is to be understood as I have been labouring to explain, in terms of two interacting 'if's of different forces. I want to claim that this is really a very simple idea, if one looks at the restrictions on branchings of worlds that it expresses; it only seems complicated because I am trying to express it in terms of 'if'.

This by itself gives us a reason to be interested in conditionals of the form 'if p then if q then r'. But there is another reason. If one was trying to convert this account of 'would cause' into an account of 'causes', one would immediately trip over the fact that much of what I have said is not of much use when C, e_1, and e_2 all actually obtain. So one would have to displace oneself from actuality, and say as if from another possible world 'if this had happened, then e_1 would have caused e_2', and the displacement would inevitably take the form of yet another antecedent. One would say, if e_1 had not occurred then it would be the case that it would have caused e_2 if it had occurred.

This suggests that we look at 'if p then if q then r', where p is incompatible with r. And in any case, whether or not a definition of 'causes' may be expected, conditionals of this kind are very puzzling. Consider if p then if not p then q'. One might imagine it to be quite straightforward: 'if p'—so we look at the nearest worlds at which p is true—'then if not p'—so we look at the worlds nearest to *those* worlds in which not p is true—'then q'—so we see if q holds at these last worlds. It is not raining; if it were raining the pavement would be wet, and that is the only way it could become wet—there are no hoses or lakes or animals around; so if it were not raining the pavement would be dry, and so under any naturally attainable conditions in which it *is* raining it is true that if it were not raining the pavement would be dry. And so one would imagine it to be true that if it were raining then if it were not raining the pavement would be dry.

But: this conditional does not even make sense, intuitively; one can't get one's understanding around it. The comprehension balks a bit at others like it, too, such as 'if it were raining then if

the pavement were dry it would not be raining'. If one understands them just as having the form 'if p then (if q then r)', though, where both 'if's are Stalnaker's conditionals, there's no difficulty; they're perfectly clear.

There's an explanation for this, if what has been said so far is in outline correct. We saw that there are many uses for 'triadic' conditionals 'if p then if q then r', where the two 'if's are not independent; the sense of the inside one is strengthened because of the effect of the outside one. I conjecture that one (nearly) always understands iterated conditionals in this way. One supposes that there is one central Stalnakerian 'nearest world in which' departure from actuality involved, and one adjusts the senses of the other 'if's to fit it. There are two easy ways of doing the adjusting. Further conditions can simply be amendments to an initial one, so that one reads 'if p then if q then r' as 'if p and q then r' (they are equivalent for a material conditional and not for a subjunctive one), and in this case 'if p then if not p then r' will obviously be very puzzling, since we haven't much use for 'if p and not p then r'. Or, the second way of adjusting, one can take the inside 'if' to be related to the outside one in the way that I called causal. One will only do this if there is some hint in the syntax that this is the intended reading, but when one does so incompatible antecedents become much less disturbing. 'If it rains then if the pavement is dry it will be because Helen managed to get her pavement-drying machine to work' is quite compatible with 'if it rains the pavement will be wet' (because Helen just can't get the machine to work, but some semi-miraculous chance might remedy its faults, and then the pavement would be dried).

Conclusions about Cause I have already stated all the conclusion I have any right to draw about 'would cause', and the effect of it all is a sort of backhand tribute to the power of the Stalnaker-Lewis conditional. For in a number of situations in which it did not seem to represent the sense that an English 'if' of a subjunctive variety had, we found that by adjusting the one really elastic factor in the analysis, the notion of the nearness of one possible world to another, things could be made to come out right. The manner of these adjustments makes a linguistic conjecture very attractive. When an 'if' is found in the scope of another 'if', or

in a context in which there are other implicit conditions (and that means therefore: always), the senses of the various conditionals, the measures of departure from actuality that they presuppose, are according to some deep and subtle and pervasive rules adjusted to one another, so that the whole combination represents an easily digestible suggestion of what might happen if. One way in which we indicate how this adjustment is to be understood is by the use of causal idioms; causes, because, it would be because of, as a result. . . . No doubt they are all different.

If this is right then 'would cause' is in a way more basic than 'causes'. For when one says what *would* cause something one is saying that it would occur if . . ., and using 'cause' to be precise about what one intends by 'if'. But when one says what *does* cause something one is, I believe, saying something extraordinarily subtle and context-dependent about the conditions under which what did happen would still have happened, if it had not happened. But thinking my way through this, or even coming to be really sure that it is right, is at the moment far beyond me.*

NOTE

* Some of the material in this paper, especially the discussion of conditionals and branching time, is developed from a draft of a longer paper by Fabrizio Mondadori and me. My debt to David Lewis—see especially 'Causation', *Journal of Philosophy* 70 (1973) 556–67, and *Counterfactuals* (Blackwells, 1973)—and J. L. Mackie—see especially 'Causes and Conditions', *American Philosophical Quarterly* 2.4 (1965), 245–55, 261–4—should be obvious. For Stalnaker's account of conditionals see his 'A Theory of Conditionals' in *Studies in Logical Theory* ed. Rescher (Blackwells, 1968). All these articles are reprinted in Ernest Sosa's *Causation and Conditionals* (Oxford U.P., 1975).