THE GOOD, THE BAD AND THE FAR-FETCHED

To judge the suitability of courses of action, the adequacy of explanations and the worthiness of acts we sometimes must evaluate hypothetical arguments. What separates good ones from bad? Take for instance arguments that string together counterfactual conditional sentences. From:

(1) If I were to turn the ignition key, I would engage the starter motor, and:

(2) If I were to engage the starter motor, the engine would start.

we might infer:

(3) If I were to turn the ignition key, the engine would start.

This argument seems about as good as can be. But what makes it good? Apparently none of its formal properties does, since other arguments seem to be both formally identical and no good at all. For instance, if someone asserts:

(4) If J. Edgar Hoover had been a Communist, he would have been a traitor to the U.S.A.

we cannot put this next to the likelihood that:

(5) If J. Edgar Hoover had been born a Russian, he would have been a Communist.

and infer:

(6) If J. Edgar Hoover had been born a Russian, he would have been a traitor to the U.S.A.

Let > be a counterfactual conditional connective, joining say 'I will turn the ignition key' and 'I will engage the starter motor' to make (1) above. Both arguments are instances of the scheme *A*>*B*, *B*>*C*/*A*>*C*, *transitivity*, and there are just two alternatives. Either transitivity is valid and we must explain what is bad about some instances, or else it is not valid and we must explain what is good about others.

In standard treatments of counterfactuals transitivity is *not* valid. The idea underlying these treatments is that *A*>*B* is to be evaluated positively just in case the consequent *B* will be, after making such adjustments as are needed to accommodate the antecedent, *A*.[1] Counterexamples to transitivity can arise when the adjustments needed to accommodate *A* are greater than those needed to accommodate *B*. Then it can happen that although *C* will be evaluated positively after making the lesser adjustments that will suffice to accommodate *B*, *C* will not be evaluated positively after making the greater adjustments needed to accommodate *A*, although *B* will be evaluated positively then as well.

[1] Both "metalinguistic" and modal treatments are based on this idea. See Nelson Goodman, "The problem of counterfactual conditionals," *The Journal of Philosophy*, 44 (1947): 113-128; see Robert Stalnaker, "A theory of conditionals," in Nicholas Resher (*ed*.), *Studies in Logical Theory* (Oxford: Blackwell, 1968); and see David Lewis, *Counterfactuals* (Cambridge Mass.: Harvard University Press, 1973).

My purpose here is to explain, assuming a standard treatment of counterfactuals, why it is that some transitive arguments are compelling but others are not. After that I shall compare my explanation with an alternative. Finally, I shall consider the prospects for a non-standard treatment of counterfactuals, in which transitivity is valid and seeming counterexamples are to be explained away.

***

In fact the engine runs. But there are things that would keep it from running. For instance, a failure of the fuel pump would. A sentence *C* & ~(*A*>*C*) expresses that *C* is the case, but were *A* the case *C* might not be the case. One such sentence that we can accept is:

> (7) The engine runs, but it is not the case that if the fuel pump were to fail then the engine would run.

A failure of the fuel pump is something that we *know* would keep the engine from running, and there are a few other things that we know would too. But there is no end of other things of which we can just *imagine* that they might have the same effect. For instance, the stare of a black cat just might. There are circumstances that just might be our circumstances, under which it is not the case that if a black cat were staring the engine would run. Imagine say a complicated unsuspected chain of natural causes and effects linking stares of cats to failures of the fuel pump,[2] or imagine something even more far-fetched, an unknown occult connection.

We can imagine such far-fetched possibilities but we do not seriously reckon with them. Instead we assume that very many things would not keep the engine from running: the stares of a black cat, repainting the barn, having lunch - most things, in fact. The negation of the earlier expression '$\underline{C}$ is the case, but were $\underline{A}$ the case $\underline{C}$ might not be' comes to: $\underline{C} \to \underline{A}{>}\underline{C}$, for example:

(8) If the engine runs, then (even) if a black cat were staring it would run.

This sentence expresses an assumption that we can ordinarily make.

One might suppose that (8) paraphrases the assumption that the staring of a black cat would not keep the engine from running. It does not, quite. For that (8) is too weak: 'the engine does not run' entails (8), and so, in a standard treatment, does 'a black cat is staring'; but intuitively neither sentence tells us anything about the effects of the staring of a black cat.[3] A better paraphrase is:

[2] For some ideas about how these things might be linked up see http://www.rube-goldberg.com, or Peter C. Marzio, Rube Goldberg, his life and work (New York: Harper and Rowe, 1973).

[3] The second entailment follows directly from the validity of $\underline{A}\&\underline{C} \to \underline{A}{>}\underline{C}$ in a standard treatment. I thank Jonathan Lowe for pointing out to me that (8) is too weak to serve as a paraphrase.

(9) Necessarily, if the engine runs then (even) if a black cat were staring it would run.

Here, the modifier 'necessarily' says that its scope holds independently of particular matters of fact: (9) holds true, under any given circumstances of evaluation, just in case (8) holds true under all relevantly similar circumstances. To qualify as relevantly similar, circumstances must be like the circumstances of evaluation in regard to law-like relations, including any causal chains or occult connections linking stares of cats to failures of fuel pumps, but may differ from them in regard to the manifest behavior of the engine and cats.[4]

Neither 'the engine does not run' nor 'a black cat is staring' entails (9). To see why, consider circumstances of evaluation in which a black cat stares straight at the car and, as a result of an unsuspected causal mechanism, the fuel pump fails and the engine does not run. Under such circumstances 'the engine does not run' and 'a black cat is staring' are both true. But (9) is false, because there are relevantly similar circumstances under which (8) is false. Under these other circumstances the engine runs very nicely, but it does so only because all cats happen to be looking the other way. The same unsuspected causal mechanism is in place, so it is not the case that if a black cat were staring, the engine would run.

[4] Compare Robert Nozick's discussion of conditionals in Philosophical Explanations (Cambridge Mass.: Harvard University Press, 1981), p. 176

Importantly, though, (9) entails (8), since any circumstances of evaluation are relevantly similar to themselves. Ordinarily we will be willing to assume that the stare of a black cat would not interfere with the running of the engine. And so we ought to be willing to accept (8), a logical consequence of this assumption.

Now I propose that it is our willingness to suppose that things will not interfere with each other, held in check by prior presumptions, which makes some transitive inferences compelling but not others. Let prior presumptions be represented by a *corpus*, a set of sentences. We can evaluate a given inference, relative to a corpus, using the following three-step procedure (P):

> *First*, for the sake of the argument add the premises to the corpus.
>
> *Second*, add as many further assumptions of the form: *C* → *A*>*C* as you can, short of introducing inconsistency (there might be several ways to do this).
>
> *Third*, see whether each result of the previous step entails the conclusion of the inference in question. If so this inference, relative to this corpus, is *good*; otherwise, it is *bad*.[5]

[5] The idea is to take some sentences, and some others, and then to add as many of the second lot as you can to the first lot, while stopping short of making the result inconsistent. This idea has appeared elsewhere in various guises: in Nelson Goodman *op. cit*., in Gerald Gazdar, *Pragmatics: Implicature, Presupposition*

According to this procedure, intuitively plausible inferences are good, relative to prior presumptions. Take the inference from (1) and (2) to (3). The salient prior presumptions are, say, what we all know about car engines. Now consider:

> (10) If the engine would start were I to engage the starter motor, then (even) if I were to turn the ignition key, the engine would start were I to engage the starter motor.

This sentence, like (8), is of the form $\underline{C} \rightarrow \underline{A}>\underline{C}$. And intuitively speaking it, like (8), is compatible with what we know about car engines together with (1) and (2), the assumptions made for the sake of the argument. For (10) is a logical consequence of the very plausible assumption that turning the ignition key would not keep it from being the case that if I were to engage the starter motor, the engine would start. Therefore (10) will be in some results of the second step of procedure (P). We might expect (10) to be in *all* of these results; suppose, for now, that it will. Then the inference from (1) and (2) to (3) is good, relative to

and Logical Form (New York: Academic Press, 1979), in Carlos Alchourron, David Makinson and Peter Gardenfors, "On the logic of theory change: partial meet contraction functions and their associated revision functions," *Journal of Symbolic Logic*, 50 (1985): 510-530, and in Ray Reiter, "A logic for default reasoning," *Artificial Intelligence*, 13 (1980): 81-132.

what we know about car engines. For (10) *completes* the inference in the sense that (3) is entailed by premises (1), (2) and (10).[6]

We can be more specific. We can choose some particular treatment of counterfactuals and corpus, and demonstrate consistency claims showing that (10) finds its way into each result of the second step of procedure (P). But there is no need for this here. That (10) ought to end up in each of them is so very plausible that we can say that in any *acceptable* treatment of counterfactuals it does. In any acceptable standard treatment of counterfactuals the inference from (1) and (2) to (3), relative to what we know about car engines, is good.

The formally identical but implausible inference from (4) and (5) to (6), on the other hand, is bad. The sentence corresponding to (10) that would complete this inference is:

> (11) If J. Edgar Hoover would have been a traitor to the U.S.A. had he been a Communist, then (even) if he had been

[6] Schematically: $A>B$, $B>C$, $B>C \rightarrow A>(B>C)/ A>C$. This inference is valid in any standard treatment. By the second and third premises, minimal adjustments to accommodate A lead us to positively evaluate $B>C$, which is to say that any further minimal adjustments needed to accommodate B will lead us to positively evaluate C. By the first premise, having accommodated A, no further adjustments are needed to accommodate B; the minimal adjustments needed to accommodate B are therefore no adjustments at all. So, having made minimal adjustments to accommodate A, we will positively evaluate C.

> born a Russian, he would have been a traitor to the U.S.A. had he been a Communist.

This sentence is incompatible with prior presumptions together with (4) and (5), for there is a prior presumption that:

> (12) It is not the case that if J. Edgar Roosevelt had been born a Russian, he would have been a traitor to the U.S.A. had he been a Communist.

Since (11) is inconsistent with (4) and (12), (11) will not be added in the second step of procedure (P). In fact it is easy to see that *no* result of the second step of (P) will entail (6). For whereas each such result is consistent (assuming, as we shall, that (4) and (5) are consistent with the salient presumptions), (6) is inconsistent with (5) and (12).[7]

***

David Lewis notes that the inference scheme: $A>B$, $(A\&B)>C/A>C$ is valid in his theory of counterfactuals.[8] This scheme is a close relative of transitivity and one might think that its validity explains the fact that some transitive inferences are compelling. Frank Jackson seems to think so. He suggests that we can "construe" compelling transitive inferences as instantiations of

[7] In a standard treatment, as can easily be seen, $A>B$ and $A>C$ entail $A>(B>C)$, so (5) and (6) entail the negation of (12).

[8] See David Lewis, *op. cit*. page 35.

this scheme.[9] Indeed, in the argument from (1) and (2) to (3), one really is inclined to understand (2) in such a way that it can be replaced by:

> (13) If I were to turn the ignition key and to engage the starter motor, the engine would start.

In the argument from (4) and (5) to (6), on the other hand, one is unwilling to replace (4) by:

> (14) If J. Edgar Hoover had been born a Russian and had been a Communist, he would have been a traitor to the U.S.A.

So it might seem plausible to say that an instance of transitivity is good if, having accepted for the sake of the argument the premises $A>B$ and $B>C$, one is willing to "strengthen the antecedent" of the second premise, replacing it by $(A\&B)>C$.

Perhaps this is what Jackson had in mind. However this may be, such a proposal is unsatisfying in the same way that it would be unsatisfying simply to say that an argument is good if, having accepted the premises, one is willing to accept the conclusion. In fact, these two proposals come to the same thing. In standard treatments the scheme:

$$A>B \;/\; (A\&B)>C \leftrightarrow A>C$$

is valid so, having accepted the premises $A>B$ and $B>C$ of a transitive argument, one ought to be willing to replace the second premise by $(A\&B)>C$ if and only if he is willing to infer

[9] See Frank Jackson, *Conditionals* (Oxford: Basil Blackwell, 1987) page 82.

the conclusion $A>C$.[10] These equivalent proposals are unsatisfying because they leave two questions unanswered. Firstly, why are we *ever* willing both to strengthen the antecedent and to infer the conclusion? And secondly, why are we willing to do both in some cases but in other cases willing to do neither?

I have answered these two questions. It is our disposition to assume that things will not interfere with each other that explains our willingness to strengthen antecedents and to draw conclusions. And it is because the exercise of this disposition is constrained by prior presumptions that we treat formally identical inferences differently. When these presumptions allow it we exercise this disposition both by strengthening the antecedent and by drawing the relevant conclusion. When they do not allow it we do neither.

***

What are the prospects for a non-standard treatment of counterfactuals that validates transitivity? With such a treatment we shall not have to explain the plausibility of arguments such as that from (1) and (2) to (3). Instead we shall have to explain the implausibility of apparent counterexamples, such as that from (4) and (5) to (6). Jonathan Lowe has taken on

[10] The truth conditions of Stalnaker *op. cit*. and those of Lewis *op. cit*. validate this scheme.

this task.[11] He explains that an argument like this second one involves something like a fallacy of equivocation: although there is a conversational context in which the first premise is acceptable, he argues, and another in which the second premise is acceptable, there is no context in which _both_ premises are acceptable.

Whatever the merits of this explanation in this particular case, it appears that other apparent counterexamples to transitivity cannot be dismissed in the same way. Here for example is a single context in which both (1) and (2) are acceptable but (3) is not. I am explaining why I might start the car without using the ignition key, by "hotwiring" the starter motor directly to the battery:

> Ordinarily, of course, _if I were to turn the ignition key the engine would start_, but not now. There is nothing wrong with the starter motor itself. _If I were to turn the key I would engage it_ as usual. The problem is with the electrical system of the car: turning the ignition key would cause the fuel pump to fail. For this reason, if I were to engage the starter motor it would be by hotwiring it, to spare the fuel pump. So _if I were to engage the starter motor the engine would start_.

To discount this apparent counterexample it will be necessary either to argue that really the premises are false in this

[11] See E.J. Lowe, "Conditionals, context and transitivity," _Analysis_, 50 (1990): 80-87; page 81.

context or the conclusion true, or that really there are more contexts of evaluation here than one. Neither alternative seems promising.

Why should we even consider a treatment that validates transitivity? Motivation is supposed to come from cases such as this. Suppose an apparently sane man asserts, so to speak in one breath, both:

(15) If an avalanche had then been taking place, there would have been snow in the valley yesterday, and:

(16) If there had been snow in the valley yesterday, I would have gone skiing.

Clearly, as Crispin Wright points out, you would be bound to wonder whether the speaker had meant to hint that he had been depressed, or some such thing.[12] There might seem to be a problem here for standard treatments of counterfactuals. On a standard treatment it seems that it cannot be the force of the inference to:

(17) If an avalanche had been taking place yesterday, the speaker would have gone skiing.

that compels us to wonder. For on a standard treatment it seems that the inference from (15) and (16) to (17) has no force at all: it is not valid and, in the relevant context, it is not good either, since the speaker's apparent sanity makes for a prior

[12] See Crispin Wright, "Keeping track of Nozick," Analysis, 43 (1983): 134-140; page 138.

presumption that (17) is false. The alleged problem is to explain the puzzlement that we are likely to feel in such a case.

But there is no real problem here for standard treatments of counterfactuals. We can explain the puzzlement in the same way that we explain it in analogous cases where counterfactuals are not even in play, and where the corresponding inferences certainly are not valid. Suppose there is a strong presumption, common to everyone involved in a normal conversation, that Minos is a man of integrity. Now someone asserts:

(18) You know, Minos is a Cretan and most of them are liars.

Then this speaker too can naturally be interpreted as having said something a bit puzzling. To paraphrase Wright, you would be bound to wonder - wouldn't you? - whether he means to hint that Minos is a liar or some such thing.

Of course the explanation is not that the inference from (18) to:

(19) Minos is a liar.

is valid. I suggest an explanation along the following lines. Typically, when a speaker asserts the premises of a readily recognizable inference scheme, together in close proximity, he does so with the intention that the hearer will recognize the scheme he has in mind and will draw the relevant conclusion. The hearer, for his part, typically recognizes both the intention and the scheme and draws this conclusion. In this particular case, though, the conclusion to be drawn, using the apparently intended scheme, is (19). Since you suppose there is a common presumption that (19) is false, you wonder just what the speaker is getting at. Surely he cannot mean that Minos is a liar - can he?

We can explain in the same way any puzzlement arising when someone asserts both (15) and (16). This time the apparently intended inference scheme is transitivity. The conclusion you would draw, using this scheme, is (17). With the apparent sanity of the speaker, though, you suppose that there is a common presumption that (17) is false, and so of course you are bound to wonder. The speaker cannot mean that he would have gone skiing in an avalanche - can he?

Notice that no puzzlement arises in a context where the speaker asserts just (16). Then there is no hint that the speaker has been depressed or any such thing, although (15) is a matter of common knowledge. The reason is that the speaker has not done anything to suggest that anyone is supposed to put (15) and (16) together to infer (17). Nor will there be puzzlement if I assert both (1) and (2) while explicitly denying (3), in the context of explaining why I propose to "hotwire" the car. In that explanatory context there is nothing to suggest that anyone is supposed to put (1) and (2) together to infer (3).

Michael Morreau,
University of Maryland at College Park