Counterfactuals

When the truth functional material conditional '$\rightarrow$ ' (or '$\supset$') is introduced, it is normally glossed with the English expression 'If ..., then ...'. However, if this is the correct gloss there are a number of surprising features. Firstly, a sentence of the form '$p \rightarrow q$' will always be true when the antecedent, p, is false; and, secondly, it will always be true if the consequent, q, is true. But there are certainly some uses of 'If ..., then ...' which do not have these features.

First consider

(1) If Bush had not won the last election, then Nader would have won it.

The antecedent of this sentence is false: Bush did win the last election. But we still don't want to say that the sentence is true. If Bush hadn't won the last election, Gore would almost certainly have done so. There was virtually no chance of Nader winning. So we don't want to read the 'If ..., then ...' as a material conditional.

Now consider

(2) If Bush had polled only twenty votes across the whole country, then he would have won the last election.

This time the consequent is true. But again we don't want to say that the conditional is true: if Bush had polled only twenty votes they would not have won the election (at least, one hopes that's true). So once again we don't want to read the 'If ..., then ...' as a material conditional.

What should we conclude? One possibility would be to say that the material conditional is just the wrong reading for the 'If..., then...' construction in English. But there are plenty of cases in which it seems to get it right. More plausible is the idea that there are two different English constructions that make use of 'If ..., then ...'; and indeed, the syntax of English bears this out. Consider the two sentences

(3) If Oswald didn't shoot Kennedy, then someone else did.

(4) If Oswald hadn't shot Kennedy, then someone else would have

Clearly these don't mean the same thing. The first is not implausibly read as the material conditional. All that is ruled out is the possibility that the antecedent is true (i.e. Oswald didn't shoot Kennedy) and the consequent is false (i.e. nobody else shot him either). But the second sentence cannot be read as a material conditional. The fact that the antecedent is false (since, let us suppose, Oswald did shoot Kennedy) doesn't, by itself, make the sentence true. So it looks as though there are two quite different 'If..., then' constructions in English, marked by the different mood of the verbs involved. In (3) the verbs are in the simple indicative mood; in (4) they are subjunctive, as indeed they are in (1) and (2) ('had shot', 'would have shot', 'had won', 'would have won' etc.).

Following the standard practice of grammarians, we'll call such conditionals 'counterfactuals', and symbolize them:

> $(P \;\Box\!\!\rightarrow Q)$

Truth Conditions for Counterfactuals

In developing truth conditiions for counterfactuals we follow the account given by David Lewis, who says (roughly):

> $(P \;\Box\!\!\rightarrow Q)$ is true iff the closest possible world (i.e. closest to the actual world) in which the antecedent, P, is true, is a world in which the consequent, Q, is also true (or, in other words, $(P \;\Box\!\!\rightarrow Q)$ is true iff the closest P-world is a Q-world).

What do we mean here by 'closest'? This is a measure of similarity. The closest P-world to the actual world is the world in which P is true which is most similar to the actual world. So the account of counterfactuals amounts to this: a counterfactual $(P \;\Box\!\!\rightarrow Q)$ is true just in case the world most similar to the actual world in which P is true is a world in which Q is true. This means in order to assess the truth value of a counterfactual we have to make an assessment about similarities between worlds; and that is going to be a rather vague business. But we shouldn't let that put us off the account. The truth value of counterfactuals is itself vague; the account should mirror that vagueness.

(Note: we said that this account was roughly that given by Lewis; in fact we have simplified his account in a number of ways. The most significant concerns our talk of *the* closest P-world. There are two ways in which there might fail to be such a world, and yet the counterfactual still be true. First, there might two or more P-worlds that are equally close; provided that these worlds are all Q-worlds, that shouldn't make the counterfactual come out false. Second, there might be an infinite series of P-worlds, each one of which is closer to the actual world than the one before—compare the infinite series of fractions 1/2, 1/4, 1/8, 1/16 … each of which is closer to zero that the one that comes before; again, provided that these are all Q-world, the counterfactual Lewis avoids these problems by saying that $(P \;\Box\!\!\rightarrow Q)$ will be true iff there is a possible world, w, which is both a P-world and a Q-world, and that any P-world which is as close or closer to the actual world than w is also a Q-world. But it's not so easy to get one's mind around this formulation; so we'll stick with our simpler approximation.)

No other world can be as similar to a world as that world is to itself. Identity is the limit case of similarity. But if that is so, then, if the actual world is a P-world, $(P \;\Box\!\!\rightarrow Q)$ will be true just in case the actual world is a Q-world. That might seem to be wrong: surely we would never say 'If Oswald hadn't shot Kennedy, someone else would have' if we knew that in fact Oswald hadn't shot him. But, as ever in providing a semantics for natural language, we need to distinguish that which is false from that which is pragmatically unacceptable on other grounds. It is true that we would normally not utter a counterfactual if we knew that its antecedent was true; but that could be because, in such circumstances, we would be in a position to assert the consequent itself, and so it would be misleading to assert something weaker. You wouldn't say 'If they were to find out, you'd be in big trouble' if knew they had found out; you'd just say: 'They've found out. You're in big trouble!' This doesn't show that the counterfactual would be false. Indeed there are good reasons for thinking that it would not be. Consider this exchange:

> A: If they were to find out, then you'd be in big trouble
>
> B: Damn! I've already told them!

Here B doesn't deny what A says, on the grounds that it's a counterfactual whose antecedent is true. Quite the reverse: B uses A's counterfactual to reach the conclusion that he is in trouble. So it seems reasonable to assume that the Lewis account is right: counterfactuals with true antecedents are true just in case their consequents are true. The reason that we don't typically assert them is pragmatic.

If this is right, counterfactuals are badly named, since they don't require that the antecedent be contrary to fact. Partly because of this, they are sometimes called 'subjunctive conditionals'. But we shall go on with the shorter name.

Counterfactual Fallacies

There are a number of valid inference patterns associated with the material conditional which are not valid for the counterfactual. We shall examine the three most important.

1. Strengthening the Antecedent

The material conditional permits strengthening of the antecedent, in the sense that all arguments of the form

$(P \rightarrow Q)$

Therefore $((P \wedge R) \rightarrow Q)$

are valid.

The same is not true of counterfactuals. Consider the argument

> If the Labour Party had not won the last election, then the Conservative Party would have won it.
>
> Therefore, if the Labour Party had not won the last election and the Communist Party had got ninety per cent of the popular vote, then the Conservative Party would have won the last election.

That is clearly not a good argument. If the Communist Party had got ninety per cent of the popular vote, they would have won the election. The Lewis account of counterfactuals explains this fact. The truth of $(P \,\Box\!\!\rightarrow Q)$ requires that the nearest P-world be a Q-world; but the nearest P-world might not be an R-world. To find the nearest world that is P and R we might have to move to a still more distant world. And that world might not be a Q-world.

2. Transitivity

The material conditional is transitive, in the sense that the following inference pattern is valid:

> $(P \rightarrow Q)$
>
> $(Q \rightarrow R)$
>
> Therefore $(P \rightarrow R)$

In contrast counterfactuals are not transitive. Consider the argument

> If J Edgar Hoover had been born a Russian, then he would have been a communist.
>
> If J Edgar Hoover had been a communist, then he would have been a traitor.
>
> Therefore, if J Edgar Hoover had been born a Russian, then he would have been a traitor.

Again that's not a good argument: if Hoover had been born a Russian he would have been a patrotic communist. Again the Lewis account of counterfactuals explains why not. $(P \Box\!\!\rightarrow Q)$ requires that the nearest P-world be a Q-world; and $(Q \Box\!\!\rightarrow R)$ requires that the nearest Q-world be an R-world. But it is consistent with both of those facts that the nearest Q-world is closer than the nearest P-world. And if that is so, the nearest P-world might fail to be an R-world.

3. Contraposition

As a final example of a counterfactual fallacy, consider the inference pattern:

> $(P \rightarrow Q)$
>
> Therefore $(\neg Q \rightarrow \neg P)$

This is valid. and so this claim is true. But once again the same does not hold for counterfactuals. Consider:

> If Boris had moved into the house, then Olga would not have moved out.
>
> Therefore, if Olga had moved out of the house, then Boris would not have moved in.

That argument is not valid. We can easily describe a state of affairs that makes the conclusion true and the conclusion false. Suppose that Olga wanted to live in the same house as Boris, but the sentiment was not reciprocated. Had Boris moved into the house in which Olga was living, Olga would have been delighted and would have stayed on. (The premise is true.) However, the house itself was a very nice one: Boris wanted to move into it, and was only put off doing so by Olga's presence. (The conclusion is false.)

The Lewis account explains why counterfactuals don't contrapose, that is, why $(P \Box\!\!\rightarrow Q)$ doesn't entail $(\neg Q \Box\!\!\rightarrow \neg P)$. $(P \Box\!\!\rightarrow Q)$ requires that the nearest P-world be a Q-world. If the nearest Q-world were nearer than the nearest P-world, then it would follow that $(\neg Q \Box\!\!\rightarrow \neg P)$. But it could be that the nearest ¬Q-world is further away still (i.e. further away than the nearest P-world). But then it would not follow that such a world must be a ¬P-world.

Different similarity measures

We mentioned above the fact that similarity measure are vague. We conclude with a brief mention of some examples which seem to show that different counterfactual sentences might require different conceptions of which worlds are more similar to the actual world. Thus consider the sentences:

> If Boston were in Florida, then Boston would be in the South.
>
> If Florida included Boston, then part of Florida would be in the North.

Both of these sentences seem to be true. If so, it seems that in assessing the first we imagine a world in which we keep Florida's borders where they are, and move Boston within them; this is the closest world in which the antecedent is true In assessing the second we imagine leaving Boston where it is, and moving Florida's borders to include it; this is closest world in which the antecedent is true. Yet it might look as though the two antecedents say the same thing. Somehow the words used indicate that they don't; different similarity measures are required. Clearly it will be no easy thing to say exactly why this happens.

Further Reading

David Lewis, *Counterfactuals* (Blackwells, 1973). This is the standard book on the subject, on which these notes are closely based.