Counterfactuals

Consider the two sentences

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

(2) If Oswald hadn't shot Kennedy, then someone else would have (had-would)

Clearly these don't mean the same thing. The first has some claim to be read as the material conditional of ordinary logic. All that is clearly 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 (1) the verbs are in the simple indicative mood; in (2) they are subjunctive ('had shot', 'would have shot').

Following fairly standard usage, we'll call sentences of the form of (1) 'indicative conditionals'; and we'll call sentences of the form of (2) 'counterfactuals' (though, since they don't always have to be counter to fact to be true, they are sometimes called 'subjunctive conditionals').

Truth Conditions for Counterfactuals

We'll symbolize counterfactuals as follows:

$(P \mathrel{\Box\!\!\rightarrow} Q)$ (this is from Lewis; an alternative, from Stalnaker, is '>')

Note that this is emphatically not the same as the strict conditional:

$\Box\ (P \rightarrow Q)$

which requires that the material condition be true in every possible world.

In developing truth conditions for counterfactuals we follow the account given independently by David Lewis and Robert Stalnaker who say (roughly):

> $(P \mathrel{\Box\!\!\rightarrow} Q)$ is true (at the actual world) 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, $(P \mathrel{\Box\!\!\rightarrow} Q)$ is true iff the closest P-world is a Q-world).

What do we mean here by 'closest'? Lewis glosses it as 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 (although even this is a technical term: for reasons we'll see when we talk about backtracking it certainly doesn't just involve comparing each atomic fact in the world (whatever those might be) and looking for the best match). So the account of counterfactuals amounts to this: a counterfactual $(P \mathrel{\Box\!\!\rightarrow} Q)$ 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.

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 you 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.

Centering

Of course it is one thing to say that a counterfactual can be true when the antecedent and consequent are both true; it is quite another to say that that is sufficient for it, that is to embrace an account in which (P $\Box\!\!\rightarrow$ Q) follows from the truth of P and Q. On the Stalnaker/Lewis account, this follows from the assumption that no world is *as similar* to the actual world as the actual world. If we remove that assumption (but keep the assumption that no world is *more similar* to the actual world than the actual world) we arrive at the hypothesis of weak centering, and (P $\Box\!\!\rightarrow$ Q) no longer follows from the truth of P and Q.

Why might we want to do this? The basic idea—proposed by Nozick, List and others—is that counterfactuals should be robust: if changing the actual world a little bit would result in the antecedent staying true whilst the consequent becomes false, that might not seem good enough for a the truth of the counterfactual. (Similarly, if the antecedent is false in the actual world, we might want to say that for the counterfactual to be true, it's not enough that the consequent be true at the closest antecedent world, but that it must be true across a band of close antecedent worlds.)

Further Reading

David Lewis, *Counterfactuals* (Blackwells, 1973).

For Stalnaker's account see his 'A Theory of Conditionals' and for a defence of the ways in which it diverges from Lewis's see his 'A Defense of Conditional Excluded Middle'.

Jonathan Bennett, *A Philosophical Guide to Conditionals* Chs. 10–21