Inverting the standard picture of counterfactual thinking

Gunnar Björnsson
Göteborg University

If the argument of this paper is correct, the semantics for counterfactual conditionals and corresponding thoughts has nothing in particular to do with whether the consequent holds in appropriate possible worlds or counterfactual scenarios. When we try both to say something substantial about how the context-relativity of counterfactuals and counterfactual thinking works, and try to explain the anatomy of "standard" interpretations of counterfactuals, it seems that counterfactual thoughts are best understood as thoughts concerned with regularities.

Here is what I take to be the prevalent picture of counterfactual thinking: When we ask ourselves whether P > Q (if P were the case, Q would be the case), we proceed by generating a representation of a counterfactual scenario containing P, and "inspect" whether it contains Q: if it does, we think that the counterfactual is true, otherwise not. Since we think that counterfactual thought can be correct or incorrect and counterfactuals true or false, we think that there is some constraint on the proper construction of this scenario. And since we think that many counterfactual truths depend on contingent facts, facts that we may or may not be aware of, the function, F, that selects the scenario does not only take as input the counterfactual postulate P, but also facts that hold in this world. Furthermore, since just about any fact in this world might be relevant for the truth of some counterfactual, the actual world, or the complete truth about the actual world, is equally a variable of the selection function. Ignoring a number of complications, the standard picture of counterfactual thinking seems to be nicely rationalized by the claim that:

> P > Q is true at a world W (in normal cases, the *actual* world) if and only if F(P, W) contains Q.

A scenario can be said to determine a set of possible worlds in which it holds: on that picture, we take P > Q to be true if and only if Q holds throughout the worlds determined by F(P, W). This, roughly, is the picture inherited from Stalnaker and Lewis.

A core task for people working within this tradition has been to identify the selection function, F: to describe how the proper counterfactual scenario is correctly generated from P and W. One thing that has emerged is that the selection function is *context-relative*. The relevant scenario for a given counterfactual conditional depend not only facts about the actual world and the antecedent, but on the context of the conditional (including the consequent).

Although context relativity is frequently acknowledged, it is less often acknowledged just how deep and widespread it is. In the first part of this paper, I remind the reader of some of the standard cases where P>Q and P>~Q both seem true once situated in the right context, and discuss what seems to be reasonable interpretations of these cases. The lesson is that the traditional picture itself tells us nothing about what it is that makes certain contents of the form *F(P, W) contains Q* contextually appropriate. What is it that gives F a certain interpretation rather than another in a given context?

This question has been addressed to *some* extent, although obliquely, in efforts to determine F for a presumed "standard" interpretation of counterfactuals. Many – such as Jackson, Barker, Bennett and Hiddleston – have argued that causation is absolutely essential to understanding counterfactuals; others – Lewis in particular – have wanted to use counterfactuals to give a reductive analysis of causation, and have consequently tried to specify F without causal notions. This difference will not be crucial here: what is important is the role that generalizations play in the truth-conditions of counterfactuals on all these accounts. Bringing out that role is the theme of the second part of the paper.

I argue elsewhere that, psychologically, counterfactuals and conditionals in general can be interpreted along two quite different lines, where one kind – what is sometimes (misleadingly) called "semifactuals" – are in effect negations of the other, primary, kind. In the third part of the paper, I bring together the first two parts, suggesting that the what we might call the conditional contents expressed by primary counterfactual conditionals are representations of facts of roughly the following form:

> CC: (A) In cases of kind C, events of kind P are always accompanied by events of kind Q, and (B) the case at issue are of kind C.

When counterfactual conditionals are used to express counterfactual thoughts, the hearer (a) identifies P guided by the "if"-clause; (b) identifies Q guided by the main clause; and (c) identifies C in part with help of the context, which will focus our attention to various features of the case at issue, and in part by going for the most salient representation of a regularity fact of the form indicated by CC(A), taking the conditional's content to be whatever content of form CC that comes first to mind when focusing on the conditional.

This suggestion makes immediate sense of some common intuitions about what counterfactual conditionals communicate. Its real strength, however, is that given an appealing account of what makes some representations of regularity facts more salient or cognitively accessible than others, it predicts not only patterns in the context sensitivity of counterfactuals, but also the complexities familiar from analysis of the standard interpretation of counterfactuals.

In the final part, I explain how this account, if correct, forces us to radically rethink the standard picture, not only with regard to how we assess counterfactual conditionals, but with regard to counterfactual thinking in general. We do not assess counterfactuals by determining whether some scenario contains Q; rather, we build such scenarios using the conditional contents of counterfactual thoughts.