Interventionism: Don't believe the hype!

996 words

The interventionist account of causation developed by Woodward in *Making Things Happen* (*MTH*) is very popular among philosophers, as evidenced by the many uses to which it has been put in recent years. Interventionist definitions of causal concepts have been used to analyze a myriad of concepts, from overdetermination to physicalism to computational explanation. They have also been put to work in attempts to solve major philosophical problems, e.g. Kim's problem of causal exclusion. In this paper I argue that, its vast popularity notwithstanding, interventionism is inadequate. The problem, I argue, stems from the modal status of interventions, and there is no easy way for interventionists to fix it.

On Woodward's view, a variable X may be a cause of another variable Y even if no intervention on X ever actually occurs. To require otherwise would result in many true causal claims, those that involve causes that have never been intervened on, being classified as false. Accordingly, Woodward admits both actual and merely possible interventions as relevant to determining the truth-values of causal claims. His definition of (type-level) direct causation, for instance, reads as follows:

> (**M**) A necessary and sufficient condition for X to be a (type-level) direct cause of Y with respect to a variable set $\mathbf{V}$ is that there be a *possible* intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in $\mathbf{V}$. (*MTH*, 59, emphasis added)

According to (**M**), and assuming a possible world semantics for possibility claims, the existence of a merely possible world in which (i) an intervention on X with respect to Y occurs, (ii) other variables in $\mathbf{V}$ are held fixed while this intervention occurs, and (iii) the occurrence of this intervention is (temporally) followed by a change in the value (or in the probability distribution over values) of Y is sufficient for X to be a direct cause of Y relative to $\mathbf{V}$. An intervention on X with respect to Y is, briefly, a manipulation that results in a change in the value of X and has an effect on the value of Y, if at all, only via its effect on the value of X.

What does Woodward mean by 'possible' in the expression 'possible intervention'? As he explains in *MTH* (132), "An intervention on X with respect to Y will be 'possible' as long as it is logically or conceptually possible for a process meeting the conditions for an intervention to occur." The problem for Woodward is that if one understands 'possible' to mean 'at least logically possible', then (**M**) does not express a sufficient condition for direct causation, and so is inadequate. Take two binary variables P and E representing, respectively, whether some individual regularly consumes birth control pills ($P = 1$ if she does, $P = 0$ otherwise) and whether this individual has epilepsy ($E = 1$ if she does, $E = 0$ otherwise). According to (**M**), P is a direct cause of E relative to $\mathbf{V}_1$: $\{P, E\}$. This is so because there is a logically possible world in which interventions on P are systematically followed by changes in the value of E. The only way for such a world not to exist

would be for the description 'An intervention on P with respect to E occurs and the occurrence of this intervention is followed by a change in the value of E' to entail a contradiction, which does not appear to be the case. As Woodward formulates it, then, (**M**) is inadequate: It does not express a sufficient condition for direct causation, since the regular consumption of birth control pills is not in any way a cause of epilepsy. The same verdict holds of other interventionist definitions, e.g. the definitions of (type-level) contributing causation and of actual causation, since their definientia appeal to the concept of direct causation.

In the remainder of the paper, I examine two objections to the argument developed above. The first consists in claiming that (**M**) does not, in fact, correctly express Woodward's intentions and that, as a result, the argument misses the target. Why suspect that this might be the case? Because Woodward talks about "the counterfactuals in (**M**)" (*MTH*, 73), and this despite the fact that (**M**) is not formulated in counterfactual terms (witness the absence of subjunctives in it). Taking the definitions Woodward sometimes uses to present his account in papers published after *MTH* – definitions that are, by contrast with (**M**), formulated in counterfactual terms – as a blueprint, I reformulate (**M**) in explicitly counterfactual terms. I show, however, that the counterfactual version of (**M**) suffers from the same defect as the original version.

The second objection is, rather, a family of objections. What unifies them is that they are all attempts to replace Woodward's requirement that interventions be at least logically possible by a stronger requirement. I show, quickly, that neither conceptual nor metaphysical nor nomological possibility can help interventionists. I also examine the possibility of requiring that interventions take place in worlds that are most similar to the actual world according to Lewis's similarity metric. I argue that though this solution *might* work, adopting it would make laws of nature an essential ingredient of interventionism when one of its 'selling points' was supposed to be its ability to account for causation and causal explanation without resorting to laws. I also argue that, if one is willing to take on laws of nature, then there is little incentive to be an interventionist given the existence of alternative accounts of causation (developed by e.g. Maudlin or Hall) that – unlike interventionism – have the quality of being of reductive. Finally, I examine the possibility of restricting the set of interventions that are relevant to determining whether X is a direct cause of Y relative to $\mathbf{V}$ by appealing to the causal model (i.e. set of structural equations and associated causal graph) for $\mathbf{V}$. I explain why doing so would render interventionist definitions of *type-level* causal concepts such as (**M**) viciously circular.