The interventionist account of causation (Woodward 2003) analyzes causal claims in terms of correlations under interventions, manipulations of variables which satisfy certain explicitly causal conditions. Campbell (2007) has recently argued that one of these conditions – the so-called ‘surgical’ constraint on interventions – runs into trouble when we try and extend interventionism to model causation in psychology. The problem, he argues, is that the constraint implies that an intervention on an intention to A should remove that intention from the influence of its rational causes, such as a belief that doing A will make one happy. But it is implausible that our interest in psychological causation is an interest in what would happen in situations in which the rational autonomy of an agent is suspended in this way.

This paper defends two claims. First, I will argue that the problem Campbell raises for interventionism is in fact an instance of a wider problem, extending beyond psychology, which I call the problem of ‘abrupt transitions’ after a similar problem for David Lewis’ counterfactual theory of causation. Second, I show how the surgical constraint is ambiguous, distinguish between a weak and strong version, and argue that the problem of abrupt transitions arises only for the strong version. I conclude that the weaker constraint is the correct constraint to adopt and explore some interesting consequences of this for causation in psychology.

I start with an introduction to the interventionist account – those familiar with it can safely skip to section 1.4.

1. Interventionism Introduced

1.1 Causal Models

Interventionism is an account of the truth conditions of causal claims which builds on a long tradition in econometrics and experimental design of using causal models as vehicles for representing causal structures. A causal model is a pair, $\mathbf{M} = \langle \mathcal{V}, \mathcal{E} \rangle$, where
\( \mathcal{V} \) is a set of variables and \( \mathcal{E} \) is a set of structural equations, one for each variable. Interventionists tend to be relatively non-committal about what exactly the variables of a causal model represent, but one option is to interpret the variables \( V_i \) of a causal model and their values \( v_i \) as representing facts (i.e. true propositions) which are related as determinables to determinates. The idea is that, in modelling the causal structure of a system, we first choose a number of distinct determinable facts about the system and represent them using variables, the different possible values of which represent different possible determinates of those determinables. For example, in modelling the causal structure of a car, we could construct a model which includes the variable \( M \), representing the fact that the car has mass. The fact that the car has mass is determined by the fact that the car has a mass of \( x \) kg, for some positive real value of \( x \) – so we let \( M \) take any positive real value to represent these different possible facts (so that \( M=m \) represents the fact that the car has mass \( m \) kg). Although it is often useful to use many-valued variables, we can also represent causal relata by means of binary variables, which take value 1 if a certain fact obtains and 0 otherwise. For example, I might represent a person by means of a model which includes the variable \( S \), such that \( S=1 \) if the person is a smoker and \( S=0 \) otherwise. 

(Are there any restrictions on choice of variables? Could we, for example, construct a model including variable \( V \), such that \( V=1 \) if it’s a full moon or Jim is eating peas or there are exactly 140 tables in the Vatican, and \( V=0 \) otherwise? Some interventionists talk as if any variable whose values represent pair-wise incompatible propositions is acceptable.\(^1\) Others, mostly implicitly, seem to appeal to some sort of naturalness constraint on causal relata, such that, for example, facts which can only be expressed in a highly disjunctive way don’t enter into causal relations.\(^2\) This is unfinished business for interventionism, but I’ll set the issue aside here.)

The structural equations of a causal model \( M = \langle \mathcal{V}, \mathcal{E} \rangle \) determine, for every variable \( V_i \) in \( \mathcal{V} \), and for every combination of values of the other variables \( V_{j\neq i} \) in \( \mathcal{V} \), what the value of \( V_i \) would be if that combination of values of the other variables were to obtain.\(^3\) For example, if our model contains just two variables, \( X \) and \( Y \), we need two

---

\(^1\) Hall cites Halpern as sympathetic to such a view; see Hall (2007: n.13)

\(^2\) Franklin-Hall (forthcoming) briefly considers such an appeal.

\(^3\) If there is no determinate fact of the matter about what the value of a variable would be if a certain combination of values of the other variables were to obtain, we move instead to a probabilistic framework – I’ll ignore this complication here.
structural equations – one which tells us, as a function of $y$, what the value of $X$ would be if the value of $Y$ were $y$, and one which tells us, as a function of $x$, what the value of $Y$ would be if the value of $X$ were $x$.

1.2 Direct Causation and Contributing Causation

Relative to a causal model, we can distinguish between direct causation and contributing causation. Direct causation is defined as follows:

\[(DC) X \text{ is a direct cause of } Y \text{ in } M \text{ iff there is a possible intervention on } X \text{ with respect to } Y, \text{ which holds fixed the values of every other variable in } M, \text{ such that, were it to occur, there would be change in the value of } Y.\]

(I haven’t said what an intervention is yet – hold tight, we’ll get there.)

We can represent causal models by means of causal graphs, in which nodes represent the variables of the model and directed edges represent direct causal relationships between them. Here’s an example:

\[\begin{align*}
\text{Z} & \rightarrow \text{Y} \\
\text{X} & \rightarrow \text{Y} \\
\end{align*}\]

Fig. 1

This represents a model in which $X$ is a direct cause of $Y$ and $Z$ is a direct cause of both $X$ and $Y$.

Call an ordered set of variables $\{V_1, V_2, ..., V_n\}$ a path from $V_i$ to $V_n$ in $M$ iff $V_i$ is a direct cause of $V_2$ in $M$, and $V_2$ is..., and $V_{n-1}$ is a direct cause of $V_n$ in $M$. The notion of a path has a handy graphical interpretation – if one can get from $V_i$ to $V_n$ by following directed edges, there is a path from $V_i$ to $V_n$.

Contributing causation is defined as follows:
(CC) \( X \) is a contributing cause of \( Y \) in \( M \) iff there is a path \( P \) from \( X \) to \( Y \), and there is a possible intervention on \( X \) with respect to \( Y \) which holds fixed the values of variables on every other path from \( X \) to \( Y \) in \( M \), such that, were it to occur, there would be a change in the value of \( Y \).

Why the insistence on holding variables on other paths fixed? Because of the possibility of so-called ‘failures of faithfulness’ (Spirtes et al 1993). Here’s an example due to Hesslow (1976). Suppose that number of birth control pills taken (\( B \)) increases the probability of contracting thrombosis (\( T \)), but also (obviously) lowers the probability of becoming pregnant (\( P \)), itself a cause of increased risk of contracting thrombosis. Here’s the graph of this model:

![Graph](image)

Now suppose also that the direct positive effect of a change in \( B \) on \( T \) is exactly counterbalanced by the negative effect of a change in \( B \) on \( T \) along the \( B \rightarrow P \rightarrow T \) path, so that the net effect of changing \( B \) on \( T \) is zero. So to ensure that \( B \) nevertheless counts as a contributing cause of \( T \) in this model, we need to intervene on \( B \) while keeping \( P \) fixed (for example, by giving varying doses of birth control pills to patients not capable of becoming pregnant).

You’ll notice that (DC) and (CC) are relativized to a model. The next step is to ‘derelativize’ these definitions to arrive at a definition of causation \textit{simpliciter}.

\[4\] Weslake (manuscript) and Woodward (2008) suggest the following definition:

\[4\] Of course, another option is to just conclude that causation is fundamentally model-relative. One finds surprisingly many comments to that effect in the causal modelling literature, although for many philosophers it represents “an objectionable loss of objectivity”, as Schaffer (2008) puts it.
(C) $X$ is a cause of $Y$ *simpliciter* iff $X$ is a contributing cause of $Y$ in some model.

Unfortunately, this can’t be exactly right, because it’s possible to construct gerrymandered models that leave out certain salient variables, relative to which intuitively unconnected variables count as causally related (see Hiddleston’s (2005) ‘bogus prevention’ cases, for example). So (C) needs to be supplemented with a restriction on what counts as an ‘apt’ or ‘appropriate’ causal model in order to be plausible (see Hitchcock (2001) for some suggestions for conditions on model aptness). This is more unfinished business for interventionism, but the complication won’t be relevant here.

1.3 Interventions

An intervention on $X$ with respect to $Y$ is characterized informally by Woodward as follows:

“Heuristically, we may think of the allowable changes in $X$ (interventions, as we have been calling them) as processes that satisfy whatever conditions must be met in an ideal experiment designed to determine whether $X$ causes $Y$” (Woodward 2003: 46).

So although every account of causation will, of course, find some connection between causal relationships and correlations of variables under experimentally idealized interventions, interventionism is unique in taking such correlations to be constitutive of causation, and not simply evidence for it (one can draw a useful analogy here with frequentism, which similarly attempts to treat long-run relative frequencies as constitutive of objective probability, and not simply evidence for it).

As will become clear, interventionism is explicitly ‘non-reductive’, in that it defines causation in terms of interventions and defines the notion of an intervention in terms of causation. The account is not viciously circular, since none of the conditions on something being an intervention on $X$ with respect to $Y$ require $X$ to be a cause of $Y$, and so the *definiens* is not included in the *definiendum*. Nevertheless, this feature of interventionism might put off those philosophers inclined to think of philosophical
analysis as always proceeding in the direction of greater fundamentality or simplicity. Interventionists are likely to reply that the proof is in the eating – if we can show how interventionism illuminates various previously obscure features of our concept of causation, then, as Strawson famously put it, “the general charge of circularity would lose its sting, for we might have moved in a wide, revealing, and illuminating circle” (Strawson 1992: 19-20).

Formally, an intervention is represented in the interventionist framework by a change in the value of an ‘intervention variable’ from its ‘off’ value to one of its ‘on’ values. Woodward suggests three conditions that a variable must satisfy in order to count as an intervention variable for $X$ with respect to $Y$. The first one is fairly obvious:

$$(I_1) I \text{ is a cause of } X.$$  

An intervention variable for $X$ should be a cause of $X$ (by which I mean cause \textit{simpliciter}).

The second condition can be stated as follows:

$$(I_2) \text{ Every path from } I \text{ to } Y \text{ (if there is one at all) goes through } X \text{ in every model containing } I, X \text{ and } Y.$$  

This is the requirement that an intervention should not affect $Y$ ‘directly’, but only, if at all, ‘though’ $X$. Otherwise, any change in $Y$ may be a consequence of the \textit{way} $X$ was changed, rather than the change in $X$ itself. Variables on paths to $Y$ that don’t go through $X$ are sometimes called ‘confounders’ and $(I_2)$ is sometimes called the constraint that interventions should ‘control for confounders’. The constraint applies to every model containing $I$, $X$ and $Y$, because interventions should control for confounders whether or not they happen to appear in our model. Successfully controlling for confounders is often a primary concern for experimental designers, and various techniques exist for doing so. The gold-standard technique is the double-blind control trial, in which subjects are assigned to either a group in which the intervention is

\footnote{Woodward also includes a fourth condition to close a loophole in $(I_2)$, but it won’t be relevant here.}
applied, or a group in which it isn’t applied, where neither the subject nor anyone else involved in the experiment knows which is which. This allows experimenters to calculate the effect of an intervention that doesn’t cause variables, such as levels of optimism in the subjects, which might be independent causes of the purported effect variable.

1.4 The Surgical Constraint

It is the third condition, however, that will be my focus in this paper. It is sometimes called the condition that interventions be ‘surgical’, in Pearl’s (2000) words. Here is how Woodward states it:

\[(I3) \text{ When } I \text{ takes one of its ‘on’ values, “} X \text{ ceases to depend on the values of the other variables that cause } X \text{ and instead depends only on the value taken by } I \text{” (Woodward 2003: 98).}\]

(Again, ‘cause’ here means cause simpliciter.)

As it stands, \((I3)\) is ambiguous, and it will important for what follows to notice why. Start by considering the following causal model (the question mark indicates the causal relation being analyzed):

\[Z \rightarrow I \rightarrow X \rightarrow ? \rightarrow Y\]

Fig. 3

In this case, what \((I3)\) requires is clear. When \(I\) takes one of its ‘on’ values, the structural equation for \(X\) in this model should be a function of the value of \(I\) only; it should not, in particular, be a function of the value of \(Z\). The motivation for this constraint is something like this. Suppose I wiggle my intervention variable \(I\) and observe no change in the value of \(Y\). This could be because \(X\) is not a cause of \(Y\). But it
could also be because changes in the value of Z cancelled out the effect of changing I, so that the net effect of changing I on X is zero – in which case it doesn’t follow from Y’s not changing, of course, that X is not a cause of Y. So in order to ensure that Y’s changing under an intervention on X is a necessary condition of X being a cause of Y, we need to ensure that when I is set to one of its ‘on’ values, X ceases to depend on the value of Z and instead depends only on the value of I.

But now consider the following causal model:

\[ I \rightarrow Z \rightarrow X \rightarrow ? \rightarrow Y \]

Fig. 4

In this model, I is not a direct cause of X. But suppose that it is nevertheless a contributing cause of X, so it satisfies \((I1)\) – it causes X ‘through’ another one of X’s causes. Suppose also that, in every model containing Z, X and Y, there is no path from Z to Y that doesn’t go through X, so that condition \((I2)\) is also satisfied. Suppose finally that both the value of Z and the value of X are completely determined by the value of I when I takes one of its ‘on’ values – the structural equations for both variables can be expressed as functions of the value of I only for this set of values. Here’s the question – does I satisfy condition \((I3)\)? On the one hand, the value of X is completely determined by the value of I. But on the other hand, the value of X continues to depend on the value of one of its other causes, namely Z – it has to, otherwise changes in I wouldn’t result in changes in X. So it’s not clear whether \((I3)\) is satisfied in such a case.

Let’s distinguish between a weak and a strong version of \((I3)\):

\[(I3)_{\text{weak}} \quad \text{When I takes one of its ‘on’ values, the value of I completely determines the value of X.}\]

\[(I3)_{\text{strong}} \quad \text{When I takes one of its ‘on’ values, the value of X ceases to depend on the values of any of its causes except I.}\]

Let’s also call a change in an intervention variable a ‘direct intervention’ if the intervention variable satisfies \((I3)_{\text{strong}}\) and an ‘indirect intervention’ if the intervention
variable satisfies \((I_3)_{\text{weak}}\) but not \((I_3)_{\text{strong}}\). \(I\) in fig.3 is a direct intervention variable, while \(I\) in fig.4 is an indirect intervention variable. Indirect interventions may, of course, involve many more intermediary variables; consider the following causal structure, for example:

\[
\begin{array}{cccccccc}
I & \rightarrow & H & \rightarrow & A & \rightarrow & X & \rightarrow & ? & \rightarrow & Y \\
& & / & / & & / & & & & \\
& & D & \rightarrow & E & \rightarrow & & & \\
& & & F & \rightarrow & G & & & \\
\end{array}
\]

So long as the value of \(X\) is completely determined by the value of \(I\), and so long as none of the variables \(A\) to \(H\) are on paths to \(Y\) that don’t go through \(X\), \(I\) counts as an indirect intervention variable for \(X\) with respect to \(Y\).

Let’s illustrate the difference between the two kinds of intervention with an example. Suppose I’m investigating the link between the level of a certain drug in a patient’s blood (\(X\)) and probability of recovery from a certain disease (\(Y\)). As it happens, a small amount of the drug (\(Z\)) is also produced naturally by the body as a by-product of a certain biological process, which varies from patient to patient. One of the causes of levels of the drug in a patient’s blood, therefore, is the amount produced naturally by that patient’s body – \(Z\) is a cause of \(X\).

Imagine two different possible manipulations of \(X\), the level of drug in a patient’s blood. In the first, a pill is given to the patient that has the effect of increasing the intensity of the biological process which naturally produces the drug in the patient’s body. The biological process is very sensitive to the dosage of the pill, so that the total level of drug in the patient’s body can be fixed quite accurately by choosing the appropriate dosage. Neither the biological process nor the pill itself, let’s suppose, are independent causes of recovery from the disease, so that \((I_2)\) is satisfied – there is no path from \(I\) to \(Y\) that doesn’t go through \(X\) in any model containing those variables. This is an indirect intervention on \(X\). It satisfies \((I_3)_{\text{weak}}\) but not \((I_3)_{\text{strong}}\). According to \((I_3)_{\text{strong}}\), the intervention should fix the total levels of the drug in the patient’s blood.
regardless of how much is being produced by the body; so in particular, in cannot change the total levels by changing the levels produced by the body.

But suppose instead that we manipulate total drug levels in the following way. The biological process producing the drug internally is allowed to operate as normal, but an instrument is set up to monitor the levels of drug produced, and to administer an additional dosage of the drug intravenously, so that the total amount of the drug in the patient’s blood is kept at a constant level, controllable by the experimenter. This is a direct intervention on $X$ – it satisfies $(I3)_{\text{strong}}$ (with some caveats to be discussed below). Variations in levels of drug produced internally are allowed to continue, but the effect of the intervention is to remove the total levels of drug in the blood from the influence of these variations.

Which is the correct constraint on interventions? The remainder of this paper will argue that $(I3)_{\text{weak}}$ is the correct constraint to adopt. Firstly, $(I3)_{\text{strong}}$ is under-motivated – there are no reasons to rule out indirect interventions, so long as the value of target variable is determinately fixed by the value of the intervention variable. And secondly, $(I3)_{\text{strong}}$ suffers from a problem, familiar from discussion of David Lewis’ (1986) counterfactual theory of causation, which I call the problem of abrupt transitions. I will discuss an instance of the problem raised by Campbell (2007) regarding the phenomenon of rational causation, causation of actions or beliefs by reasons to act or believe. Contra Campbell, however, I will argue that the problem extends beyond psychology. Campbell’s reaction is to drop the surgical constrain on interventions altogether, in certain cases – I argue instead that the correct reaction is to adopt the weaker of its two disambiguations.

2. Rational Causation

2.1 Intervening on Intentions

Campbell invites us to consider the following example. Suppose Jane believes that marrying Tim will make her happy. This belief causes her to form an intention to marry Tim, which in turn causes her to marry Tim. Let’s model the situation using the variables $\text{BELIEF}=1$ if Jane believes that marrying Tim will make her happy and 0 otherwise, $\text{INTENTION}=1$ if Jane intends to marry Tim and 0 otherwise, and $\text{MARRY}=1$ if Jane marries Tim and 0 otherwise, so that $\text{BELIEF}$ causes $\text{INTENTION}$ causes $\text{MARRY}$. According to
interventionism, the fact that \textit{INTENTION} causes \textit{MARRY} means that were an intervention on \textit{INTENTION} with respect to \textit{MARRY} to be performed, (keeping the relevant things fixed) there would be a change in \textit{MARRY}.

What is it to intervene on someone’s intentions? As Campbell points out, we are well accustomed to manipulating the intentions of others. If, for example, I didn’t want Jane to marry Tim, I might manipulate her intentions by telling her, “You think that marrying Tim will make you happy, but it won’t”. Let’s introduce a variable \( I \), such that \( I=1 \) if I say to Jane, “You think that marrying Tim will make you happy, but it won’t”, and \( I=0 \) otherwise. Plausibly, \( I \) has the following causal structure:

\[
I \rightarrow \text{BELIEF} \rightarrow \text{INTENTION} \rightarrow \text{?} \rightarrow \text{MARRY}
\]

Fig. 5

In other words, whether or not I say to Jane, “You think that marrying Tim will make you happy, but it won’t” causes whether or not Jane believes that marrying Tim will make her happy, which causes whether or not she intends to marry Tim. I am manipulating Jane’s intentions \textit{through} the usual, rational causes of her intentions. As Campbell puts it, by intervening in the way shown in fig.5, “You would be appealing to the rationality of the subject” (Campbell 2007: 61).

According to Campbell, however, this kind of manipulation doesn’t count as an intervention on Jane’s intention, because the intention continues to depend on the rational causes of her intention. This suggests that Campbell is interpreting the surgical constrain on interventions along the lines of \((I_3)^{\text{strong}}\) – interventions must act directly on target variables, removing them from the influence of any of their other causes. So, according to Campbell, an intervention on intentions which satisfies the surgical constraint should look like this:
When I takes one of its ‘on’ values, whether or not Jane forms an intention to marry should cease to depend on whether or not she believes that marriage will make her happy (represented by the dotted line in fig.6). An intervention on Jane’s intention should ensure that the intention is formed, whether or not there is any rational reason to form it.

2.2 Mental Surgery

Now there are some philosophers who would simply deny that such a direct intervention on intentions is even possible. They would deny, in other words, that it’s coherent to speak of intentions to A being manipulated independently of beliefs about A. On this form of ‘mental holism’, the content of an intention is partly constituted by the position of that intention in a broadly rational ‘web’ of mental states. Changing other mental states therefore de facto changes the content of an intention. As Campbell puts it, “The mind has to be organized in a broadly rational way, for there to be a mind there at all” (Campbell 2010: 71). There is no such thing, nor could there be such thing, as an agent who believes that marriage will make her happy, has no reason not to get married, yet nevertheless fails to intend to get married, because it is an a priori prerequisite on something exemplifying mental states at all that those mental states fit together in a broadly rational way. In a slogan: One cannot simply pick-and-mix mental states. Therefore, a manipulation of mental states which satisfies the surgical constraint, as Campbell interprets it, is a conceptual absurdity. And therefore, the counterfactual delivered by the interventionist analysis of ‘Jane’s intention to marry caused her to marry’ has an impossible antecedent.

As Campbell remarks, it’s this kind of view of the mental that “underpins some of the hesitation philosophers have felt in talking about mental causation at all” (Campbell
But Campbell himself is no friend of mental holism (see Campbell 2010). He grants that a direct intervention on intentions is coherent. His point is rather to emphasise just how strange such an intervention would be. What we are asking for is a manipulation of Jane’s intentions that disrupts the sensitivity of Jane’s intentions to rational evaluation. “It does not happen very often, if it happens at all, that a person’s rational autonomy is suspended and some alien force seizes control over whether that person has a particular intention”. Indeed:

“One who seemed to find him- or herself in that situation – someone who encountered in introspection an intention that seemed to have been the direct result of someone else’s long-standing objectives, interests, preferences, and so on – would experience this as thought insertion, the feeling that someone else’s token thought has been pushed into your mind, one of the symptoms of schizophrenia…It is exactly this situation that we are envisaging, though, when we think in terms of surgical intervention on possession of an intention” (Campbell 2007: 62).

Now, certainly, the kinds of interventions being envisaged here are not the kind of manipulations of intentions that are actually carried out by experimental psychologists and sociologists. Consider, for example, the studies reviewed by Webb and Sheeran’s (2006) recent meta-analysis of the experimental evidence that intentions are causes of behaviour. Webb and Sheeran include 47 studies in which the effects of a particular ‘intervention’ on subjects’ intentions to engage in a certain kind of behaviour – e.g. safe sex, smoking, visiting an internet site, and so on – and their subsequent behaviour is measured. The results show that the interventions had a sample-weighted average effect of size 0.66 on intentions and an average effect of size 0.36 on behaviour, showing, the authors conclude, that the correlation between intentions and behaviour is indeed causal (though much weaker than many philosophers might assume!). The kinds of interventions involved in the studies included in Webb’s and Sheeran’s analysis, however, are (at best) indirect interventions, and not direct interventions. For example, in Brubaker and Fowler (1990), college males were presented with “persuasive messages” on audiotape, in which a doctor challenges misconceptions about testicular self-examination (TSE) raised by students, before urging the listener to carry out the procedure once a month. The intervention was found to have an effect on both intentions
and behaviour. But Brubaker and Fowler explicitly acknowledge that the intervention affected the intentions by affecting the beliefs of the subjects which constitute the rational causes of the forming of their intentions: “The experimental message...was designed to alter subjects' beliefs about the outcomes of performing TSE” (Brubaker and Fowler 1990: 1414). Hence, the intervention fails to remove the target variable from the influence of its other causes, as $(I^3)_{\text{strong}}$ requires.

Of course, we shouldn’t expect the interventionist account to analyse causation in terms of the kinds of interventions that are actually performed in science. All kinds of limitations usually make the ideal experiment impossible to perform, and so we should expect a certain amount idealization in an account of the truth-conditions of causal claims in terms of the experimental procedures used to test them. (Again, the analogy with frequentism is helpful – any plausible frequentist account should analyze objective probability in terms of long-run frequencies, longer, indeed, than anyone is usually in a practical position to determine). But nevertheless, Campbell argues, “it is not credible that our interest in psychological causation is an interest in what would happen under such idealized conditions of alien control” (Cambpell 2007: 62). It’s just implausible, idealization or no idealization, that when we say that Jane’s intention to marry Tim caused her to marry Tim, we’re saying something about what would have happened in cases so far removed from our own psychological lives as to be unrecognizable.

### 3. The Problem of Abrupt Transitions

#### 3.1 Transition Periods

Campbell’s conclusion is that if interventionism is to be extended to psychology, the conditions on interventions will have to be amended. What he doesn’t realize, I think, is that the problem is not limited to psychology. Consider the following example, taken from Woodward (2003: 142-4), and apparently originally discussed by Lewis in an unpublished lecture. I’m driving in the right-hand-lane of a dual-carriageway, and realize too late, at time $t$, that I need to take the next exit. I miss the exit, and as such I am late for my meeting. Whether or not I am in the right-hand lane at $t$ causes whether or not I am late for my meeting. According to the interventionist analysis, this is so because had an intervention on whether or not I am in the right-hand lane at $t$ been performed, there would have been a change in whether or not I am late for my meeting.
So what would an intervention on whether or not I am in the right-hand lane at $t$ be like? One natural thought is something like this: someone calls me up some time before $t$ and reminds me that I need to take the next exit. But no intervention of this kind would satisfy ($I_3$)$_{\text{strong}}$. Whether or not I am in the right-hand lane at time $t-1$, after all, is a cause of whether or not I am in the right-hand lane at $t$, for any sufficiently fine-grained units of time. A direct intervention on whether or not I am in the right-hand lane at $t$ should remove this variable from the influence of my location at any earlier time. In other words, the intervention must ensure that I end up in the left-hand lane at $t$, wherever I happen to be at any arbitrarily small time before that. One can certainly imagine interventions like that – for example, we could place some kind of portal in the right-hand lane, which has the effect of instantly teleporting me into the left-hand lane if I happen to go through it. We’d probably also have to do some other things to ensure that this intervention doesn’t have any independent effect on whether or not I am late for my meeting, in violation of ($I_2$) – for example, we’d have to wipe the memories of the drivers around me, so that they don’t become too alarmed at the sudden disappearance and reappearance of my car and crash into me, thereby making me late for my meeting.

On the other hand, an indirect intervention on whether or not I am in the right-hand lane at $t$ is perfectly consistent with what Lewis calls a ‘transition period’ between the intervention being performed and its effect on the target variable being realized. Lewis famously insisted on such a transition period in his counterfactual analysis, arguing that “we should sacrifice the independence of the immediate past to provide an orderly transition from actual past to counterfactual present and future” (Lewis 1979: 463). Woodward, however, seems to explicitly rule out interventions of this kind:

“…the correct possible world to consider [is] the world in which there is no transition period and in which all of the independent effects that the intervention that places my car in the [left-hand] lane would otherwise have on my exiting are cancelled or removed…In other words, in contrast to Lewis, the interventionist account tells us that we should avoid transition periods entirely” (Woodward 2003: 144).

But notice that we can make a similar complaint to the one Campbell makes about the interventionist treatment of rational causation in this case too. Even if it’s coherent to imagine an intervention which suspends the influence of my position at $t$
from my position at earlier times, and even if the interventionist account gets the right result in this case, it still seems implausible that when I say that my being in the right-hand lane at $t$ caused my lateness, I’m saying something about what would have happened in such scenarios, involving portals and mind-wiping devices, so far removed from our everyday experience as to be virtually unrecognizable. And, of course, there is nothing special about this case either. A direct intervention on whether on not Luis Suarez scores should ensure that he scores whether or not he is substituted a minute beforehand; a direct intervention on whether a light comes on should ensure that it comes on whether or not electrons in the circuit start moving; and a direct intervention on whether or not a tree grows in my garden should ensure that it grows regardless of the presence or otherwise of a seed in the soil (or the soil itself). If $(I_3)_{\text{strong}}$ is applied at a sufficiently fine-grained level of detail, many interventions start to look really weird. Call this the problem of ‘abrupt transitions’. The problem Campbell points out for interventionism in psychology – at least if we are unmoved by mental holist worries about the independent manipulability of mental states – is just a special case of this problem.

3.2 Allowing Indirect Interventions

How should we react to the problem of abrupt transitions? In the case of causation in psychology at least, Campbell concludes that the answer is to abandon the surgical constraint entirely – i.e. we should analyze causation in terms of manipulations which only satisfy $(I_1)$ and $(I_2)$. These kinds of manipulations are sometimes called ‘soft’ interventions, and there is a burgeoning literature on their epistemological value in learning about causal structures under certain assumptions (see, for example, Eberhardt and Scheines (2006)). But, to repeat, what we’re after is not an account of causal learning, but an account of the truth conditions of causal claims, and for that purpose, something like the surgical constraint is needed – without it, a change in $Y$ brought about by an intervention on $X$ with respect to $Y$ cannot be a necessary condition of causation, since a causal connection may exist even though no change in $Y$ occurs following the intervention (because no change in $X$ occurs either).

Another option is to just refuse to be moved by the problem of abrupt transitions. Interventionism gets the right results – it doesn’t misidentify a non-causal correlation for a causal one or vice-versa – so that should be that. In support of this, it’s worth
pointing out that even $(I_3)_{\text{weak}}$ is a remarkably strong constraint, applied at a sufficient level of detail. Consider our drug example again. Suppose that as before, $X$ represents the level of drug in a patient’s blood, and $Y$ represents the probability of that patient recovering from a certain disease. This time, however, let $Z$ represent whether or not the world explodes. Clearly, whether or not the world explodes causally affects the level of the drug in a patient’s blood – i.e. $Z$ is a cause of $X$. A direct intervention on the levels of the drug in a patient’s blood, therefore, should ensure that her blood contains that level of the drug $\text{whether or not the world explodes}$. This is some intervention indeed. But since $(I_3)_{\text{weak}}$ requires the level of drug to be completely determined by the value of the intervention variable, it requires at least that the intervention alone is sufficient to ensure that the world doesn’t explode. And that’s still pretty impressive. Causal modellers are unlikely to be moved by these kinds of recherché possibilities, but nevertheless, despite Woodward’s insistence that “there will be realistic cases in which manipulations carried out by human beings will qualify as interventions”, no actual experiment, human or otherwise, is ever going to completely satisfy even $(I_3)_{\text{weak}}$. So if an intervention is an unattainable ideal, why worry about the problem of abrupt transitions?

The main problem with $(I_3)_{\text{strong}}$, however, is that it represents an ideal that experimentalists don’t seem to feel any pressure to even try and approach. Typically, an experiment will measure the effect of a change in the intervention variable on both the target variable and the purported effect variable. Experimentalists will be concerned to ensure that, as best as possible, the value of the target variable is fixed entirely by the value of the intervention variable. They won’t typically be concerned with ensuring that the intervention variable doesn’t act on the target variable ‘through’ any other variable – except if there is reason to think that that intermediate variable is a confounder.

This latter concern seems to be what Campbell and Woodward have in mind when they motivate $(I_3)_{\text{strong}}$:

“…suppose we leave intact the belief that doing X will make one happy. Then it is possible that the belief that doing X will make one happy causes both formation of the intention to do X and also directly causes performance of the action itself. In that case the intention to do X will be correlated with doing X even though the intention plays no role in causing the action” (Campbell 2007: 61-2).
Similarly, Woodward insists on eliminating transition periods between interventions and target variables, “because they may introduce factors that affect the effect independently of the putative cause” (Woodward 2003: 144-5).

But this possibility is already taken care of by \((I_2)\). If there is a path from a variable \(Z\) to \(Y\) that does not go through \(X\), then \((I_2)\) states that an intervention variable \(I\) for \(X\) with respect to \(Y\) should not also cause \(Z\). If there is no such path, however, then there is no reason why \(I\) shouldn’t cause \(Z\) as well as \(X\). In other words, there seems to be no reason why an intervention cannot act ‘through’ other variables, so long as those variables are not independent causes of the purported effect variable. Given \((I_2)\), there seems to be no reason to rule out indirect interventions.

3.3 Rational Causation Revisited

To summarize, once we distinguish between \((I_3)_{\text{weak}}\) and \((I_3)_{\text{strong}}\), it’s clear that \((I_3)_{\text{weak}}\) is the correct constraint on interventions to adopt. What does this mean for interventions on intentions? I think it’s plausible that, in most normal cases, there is no independent causal route from the rationalizing causes of an intention to do an action \(A\) to \(A\) itself that doesn’t go through an intention to do \(A\). Therefore, if we adopt \((I_3)_{\text{weak}}\), it follows that an intervention on the majority of intentions needn’t suspend the rational autonomy of an agent; and therefore, it follows that our interest in the causal consequences of intentions is not usually an interest in what would happen in cases where the rational autonomy of an agent is so suspended.

There are exceptions, however, as demonstrated by well-known problem cases in the philosophy of action literature. Suppose, for example, that Johnny is a secret agent, who forms a belief that there is poison in his wine. This causes him to form an intention to drop (and therefore smash) his glass, thereby avoiding having to drink the wine without arousing suspicion. However, whenever Johnny believes he is in danger he gets nervous, and this causes his palms to sweat which makes him more likely to drop whatever he’s holding (he never did successfully complete secret agent training). In this case, there is a path from whether or not Johnny believes that his glass is poisoned to whether or not he drops the glass, which does not go through whether or not Johnny intends to drop his glass. An intervention on Johnny’s intention to drop his glass must
therefore, given $(I_2)$, manipulate his intention without affecting whether or not he believes the wine is poisoned; i.e. without affecting the rational causes of Johnny’s intention. So in some cases, interventions of the kind Campbell criticizes, which suspend the rational autonomy of an agent, are necessary to establish a causal connection, even given $(I_3)_{\text{weak}}$. But these are special cases. In the majority of cases, according to $(I_3)_{\text{weak}}$, it is possible to intervene on intentions by manipulating the usual, rational causes of those intentions, provided that whether or not an intention is formed is determined entirely by the intervention.

I’ve argued that Campbell’s problem is an instance of a wider problem for interventionism, one which arises from an ambiguity in the standard statement of the surgical constraint on interventions. The correct solution is to adopt the weaker of the two disambiguations – there is no reason to rule out indirect interventions in an interventionist analysis of causal claims. In particular, this means that, except in certain interesting special cases, it doesn’t follow from interventionism that our interest in psychological causation is an interest in the behaviour of agents lacking rational autonomy.
References


Weslake, B. (manuscript). Exclusion excluded.
