



Interventionism and Mental Surgery

Alex Kaiserman¹

Published online: 17 September 2018
© The Author(s) 2018

Abstract

John Campbell has claimed that the interventionist account of causation must be amended if it is to be applied to causation in psychology. The problem, he argues, is that it follows from the so-called ‘surgical’ constraint that intervening on psychological states requires the suspension of the agent’s rational autonomy. In this paper, I argue that the problem Campbell identifies is in fact an instance of a wider problem for interventionism, extending beyond psychology, which I call the problem of ‘abrupt transitions’. I then defend a solution to the problem, which replaces the surgical constraint with a weaker constraint on interventions that nevertheless does all the work the surgical constraint was designed to do. I conclude by exploring some interesting consequences of this weaker constraint for causation in psychology.

1 Introduction

The interventionist account of causation is an analysis of causal claims in terms of correlations under *interventions*, manipulations of variables satisfying certain explicitly causal conditions. Campbell (2007) has claimed that one of these conditions—the so-called ‘surgical constraint’—runs into trouble when we try to apply interventionism to causation in psychology. The problem, he argues, is that it follows from the surgical constraint that an intervention on an intention should remove that intention from under the influence of its rational causes; and it’s 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 aims to do two things. 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 discussed by David Lewis. Second, I’ll argue that the problem of abrupt

✉ Alex Kaiserman
alexander.kaiserman@philosophy.ox.ac.uk

¹ Balliol College, University of Oxford, Broad Street, Oxford OX1 3BJ, UK

transitions can be solved by replacing the surgical constraint with a weaker constraint on interventions, one which nevertheless does all the work the surgical constraint was designed to do. I conclude by exploring some interesting consequences of this weaker constraint for causation in psychology.

The paper begins with an (opinionated) introduction to interventionism; readers already familiar with the view can safely skip to Sect. 3.

2 Interventionism Introduced

Barometer readings are correlated with weather patterns. When barometer readings go down, stormy conditions tend to follow; and when barometer readings go up, fair weather conditions tend to follow. That, after all, is why we use barometers to forecast the weather. But correlation does not imply causation, to use a well-worn cliché. A fall in the reading of a barometer is not a cause of the storm which follows it. How do we know this? Because we know that barometer readings and weather patterns are correlated only because they have a *common* cause, namely, the atmospheric pressure; and we also know that if we were to change the reading of a barometer in a way that doesn't also change the atmospheric pressure—by opening up the barometer and fiddling about with the needle, for example—it would make precisely no difference to the weather. It's by means of exactly these kinds of targeted manipulations,¹ informed by prior causal knowledge, that causal hypotheses are actually tested in the special sciences.

Although every account of causation will, of course, find some connection between causal relationships and correlations under experimentally idealized manipulations, interventionism is unique in taking such correlations to be *constitutive* of causation, and not simply evidence for it.² Roughly speaking, the interventionist defines actual causation, not in terms of what would have happened if the cause hadn't *occurred*, but rather in terms of what would have happened if the cause had been prevented from occurring *by means of an intervention*. The concept of an intervention is then itself defined in causal terms. As an analysis of causal claims, then, interventionism is explicitly and unashamedly non-reductive. But it nevertheless purports to be illuminating—in just the sense that functionalism about the mind seeks to better understand mental states in terms of the functional relations between them and not simply by defining each one individually in behavioural terms, interventionism seeks to better understand both causation and experimental

¹ Or, more commonly, by means of statistical techniques that allow us to *simulate* the effects of these kinds of targeted manipulations.

² This illuminating characterization of interventionism is borrowed from Franklin-Hall (2016, p. 556). One can draw a useful analogy with frequentism about probability here, which similarly attempts to treat long-run relative frequencies as constitutive of objective probability, and not simply evidence for it—see Wallace (2012, ch. 4) for helpful discussion.

manipulation in terms of the connections between them and not simply by defining each of them individually in non-causal terms.³

Woodward's (2003) formulation of interventionism belongs to a long tradition of using *causal models* as vehicles for representing causal structures.⁴ A causal model is an ordered pair $\mathbf{M} = \langle \mathbf{V}, \mathbf{E} \rangle$, where \mathbf{V} is a set of variables and \mathbf{E} is a set of 'structural equations', one for every variable in \mathbf{V} . The variables represent "properties or magnitudes that, as the name implies, are capable of taking more than one value" (Woodward 2003, p. 39). The simplest kind of variable is one which can take two values—say, 0 and 1—such that $X = 1$ if a particular event occurs and $X = 0$ otherwise. But multi-valued variables are possible too; for example, we might represent the causal structure of a car by constructing a model which includes a variable X that can take any positive real number as a value, such that $X = x$ if and only if the mass of the car is x kg.

The structural equations describe, for each variable $V \in \mathbf{V}$ and each combination of values of the other variables in \mathbf{V} , what the value of V *would* be if the other variables were assigned that combination of values *by means of interventions with respect to V*.⁵ (I haven't said what an 'intervention with respect to V ' is yet—hold tight, we'll get there.) Here's an example of a structural equation:

$$X := Y + Z \quad (1)$$

(1) says that, for every combination of values of Y and Z , were they to be assigned those values by means of interventions with respect to X , the value of X would be equal to the sum of the values of Y and Z .

Relative to a causal model, Woodward starts by defining the notion of a 'direct cause'⁶:

DIRECT CAUSATION (DC): X is a direct cause of Y in $\mathbf{M} = \langle \mathbf{V}, \mathbf{E} \rangle$ if and only if there is a possible intervention on X with respect to Y , and some combination of values of the other variables in \mathbf{V} , such that according to \mathbf{E} , were the intervention to occur while all the other variables in \mathbf{V} were held fixed at that combination of values by interventions with respect to Y , there would be a change in the value of Y .

³ See Woodward (2003, p. 106). This analogy suggests an interpretation of interventionism as a first step in a fully-fledged reductive analysis of causation along the lines of the 'Canberra Plan' (see Lewis 1970; Jackson 1998). The idea is that we would take the conjunction of all our interventionist definitions, 'Ramsify out' all the causal predicates by replacing them with higher-order variables bound by higher-order existential quantifiers, and then search for non-causal properties that witness the resulting sentence. Although interventionism "might be supplemented by any one of a number of different stories about metaphysical foundations", however, it "does not attempt to provide such foundations", according to Woodward (2008, pp. 194–5).

⁴ See especially, Pearl (2000), Spirtes et al. (2000).

⁵ There may be no determinate fact of the matter about what the value of X would be were the other variables in \mathbf{V} set to some combination of values by means of interventions with respect to X . Such systems can be alternatively represented by means of 'indeterministic' causal models, whose structural equations specify the *probability* of X taking a certain value conditional on the relevant interventions occurring. But I'll continue to assume a deterministic framework here for simplicity.

⁶ See Woodward (2003, p. 55). Here and elsewhere, I diverge slightly from Woodward's exact wording of the definitions for ease of exposition.

Why the insistence on holding other variables fixed? Because of so-called ‘failures of faithfulness’. Here’s an example due to Hesslow (1976). During pregnancy, the human body produces more blood-clotting proteins, leading to an increased risk of developing deep vein thrombosis. Oestrogen-based contraceptive pills are also known to increase the production of blood-clotting proteins. But of course, whether or not one is taking contraceptive pills (C) is a cause of one’s probability of becoming pregnant (P), which is a cause of one’s chance of developing thrombosis (T). Now suppose the negative effect of using contraceptive pills on the chance of developing thrombosis along the $C \rightarrow P \rightarrow T$ route is *exactly cancelled out* by the positive effect of using contraceptive pills on the chance of developing thrombosis along the direct $C \rightarrow T$ route, so that the net effect of changing C on the value of T is zero. Nevertheless, (DC) delivers the result that C is a direct cause of T in the model containing these three variables, because were we to intervene on whether or not the patient is using contraceptive pills *while holding fixed* her probability of becoming pregnant—by replacing the use of contraceptive pills with a different form of contraception, for example—there *would* be a change in the patient’s chance of developing thrombosis.

We can represent causal models by means of *causal graphs*. A causal graph contains a node for every variable of the model and a directed edge between two nodes for every direct causal relationship. Here, for example, is the causal graph of the model containing C , P and T in the example above:

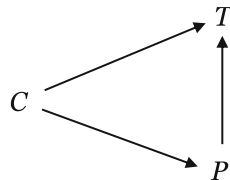


Fig. 1

We can also define the notion of a *path* between variables: an ordered n -tuple of variables $\langle V_1, V_2, \dots, V_n \rangle$ is a path from V_1 to V_n in \mathbf{M} if and only if V_1 is a direct cause of V_2 in \mathbf{M} , V_2 is a direct cause of V_3 in \mathbf{M} , and so on, up to V_n . In the causal model illustrated in Fig. 1, for example, there are two paths from C to T , one of which goes through P .

We’re now able to define a more general relation between variables,⁷ as follows:

⁷ Woodward (2003, p. 57) calls this relation “type-level causation”. This strikes me as a misnomer, however. Type-causal claims, at least as the term is generally used in the philosophy literature, are claims like ‘Smoking causes cancer’. These are *generic* claims, of the same semantic class as ‘Chickens lay eggs’ or ‘Lying is wrong’ (for an introduction to the semantics of generics, see Leslie 2012). Variable-level causal claims, however, are claims like ‘The number of cigarettes smoked by John is a cause of his chance of developing cancer’. These are *not* generic claims – rather, they predicate a relation between two token variables. See also Menzies (2008, p. 206) on this point.

VARIABLE-LEVEL CAUSATION (VC): X is a variable-level cause of Y in $\mathbf{M} = \langle \mathbf{V}, \mathbf{E} \rangle$ if and only if:

- there is a path P from X to Y in \mathbf{M} , and
- there is a possible intervention on X with respect to Y , and some combination of values of the other variables in \mathbf{V} not on P , such that according to \mathbf{E} , were the intervention to occur while all of the other variables in \mathbf{V} not on P were held fixed at that combination of values by interventions with respect to Y , there would be a change in the value of Y .

For example, in the model illustrated in Fig. 2 below, X is a variable-level cause of Y if and only if there is an intervention on X with respect to Y which results in a change in the value of Y when the value of Z (but not, of course, the value of W) is held fixed at some value.

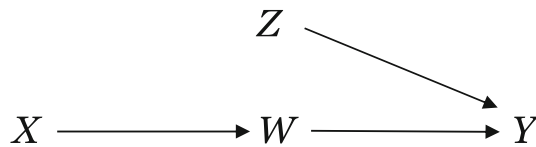


Fig. 2

Having defined a relation between variables, Woodward then defines a relation between *actual values* of variables, as follows (Woodward 2003, p. 77):

ACTUAL CAUSATION (AC): $X = x$ is an actual cause of $Y = y$ in $\mathbf{M} = \langle \mathbf{V}, \mathbf{E} \rangle$ if and only if:

- the actual values of X and Y are x and y , respectively,
- there is a path P from X to Y in \mathbf{M} , and
- there is a possible intervention on X with respect to Y such that according to \mathbf{E} , were the intervention to occur while all of the other variables in \mathbf{V} not on P were held fixed at their *actual* values by interventions with respect to Y , there would be a change in the value of Y .

You'll notice that (VC) and (AC) are both relativized to a causal model. There have been some attempts to 'de-relativize' these definitions (see Weslake forthcoming; Woodward 2008). As Hitchcock (2001) argues, such a project is likely to require some kind of restriction to 'apt' or 'appropriate' causal models.⁸ This is unfinished business for interventionism, but it won't be relevant here.

⁸ Some are happy to conclude that causation is a fundamentally model-relative concept—see Halpern and Pearl (2005, p. 845), for example.

3 Surgical Interventions

Informally, Woodward describes an intervention on X with respect to Y as a process that satisfies “whatever conditions must be met in an ideal experiment designed to determine whether X causes Y ” (Woodward 2003, p. 46). Formally, an intervention on X with respect to Y is a change in the value of an *intervention variable* for X with respect to Y from 0 (its ‘off’ value) to 1 (its ‘on’ value). Intervention variables can be defined as follows⁹:

INTERVENTION VARIABLE (IV): I is an intervention variable for X with respect to Y if and only if, in every model \mathbf{M} containing I , X and Y :

- (I1) I is a variable-level cause of X in \mathbf{M} ;
- (I2) When I takes its ‘on’ value, the value of I is the only actual cause of the value of X in \mathbf{M} ;
- (I3) Every path from I to Y in \mathbf{M} (if there is one at all) goes through X ¹⁰;
- (I4) I is statistically independent of every variable in \mathbf{M} that is on a path to Y that does not go through X .

It’s important to note that, unlike (DC), (VC) and (AC) above, (IV) is *not* relativized to a model—if a variable I is an intervention variable for X with respect to Y , it is so *simpliciter*. In particular, for a variable I to count as an intervention variable for X with respect to Y , it must satisfy (I1)–(I4) in *every* model containing those three variables, not just in the model we happen to be considering.¹¹

To motivate (IV), it will be helpful to consider an example. A high school student in the USA will on average receive better grades on standardised tests if she attends a private school than if she attends a state school (Coleman and Hoffer 1987). But whether this is a *causal* correlation is disputed. Suppose we tried to test whether private school attendance is a cause of better educational outcomes by means of the following experiment. We recruit a number of participants with children approaching high school age, who cannot afford to send their children to private schools. We then divide these participants into two groups. To the participants in one group, we give a large amount of money (enough to cover private school fees). We then compare the final grades of the children in the first group with the final grades of those in the second.

This, of course, would be a terrible experiment, for several reasons. First, we can’t be sure that increasing parental income isn’t itself an *independent* cause of better educational outcomes, for reasons that have nothing to do with private schools. Perhaps, for example, increasing parental income also has the effect of

⁹ See Woodward (2003, p. 98); I follow Weslake’s (forthcoming) formulation of this definition.

¹⁰ The parenthetical is important here—none of the conditions on interventions imply that I is an intervention variable for X with respect to Y only if X is a cause of Y . If they did, (VC) would be viciously circular (the *definiendum* would be contained in the *definiens*).

¹¹ See especially Baumgartner (2013, pp. 6–7, 2017, p. 340). Woodward himself is explicit on this point: “[A] look at [IV] makes it clear that there is no explicit or obvious relativization to a variable set... the intervention must be uncorrelated with *all* potential confounders, not just with all confounders that happen to be in some variable set” (Woodward 2008, p. 202).

relieving pressure on children to supplement household income through part-time work, freeing up more time to study. So if we *do* observe a difference in final grades between the children in the two groups, we can't be sure that this is because of a causal connection between private school attendance and educational outcomes, as opposed to an independent causal connection between parental income and educational outcomes.

(I3) is designed to address this problem. What (I3) requires is that an intervention variable for X with respect to Y should not cause Y 'directly', but only, if at all, 'through' X . Let $I = 1$ if the parent of a particular child is given the large sum of money and 0 otherwise, let $S = 1$ if the child attends private school and 0 otherwise, let G be a variable representing the child's final grades, and let $W = 1$ if the child is in part-time work and 0 otherwise. Here, plausibly, is the graph of the model that includes these four variables (the question mark indicates the causal relation being analysed):

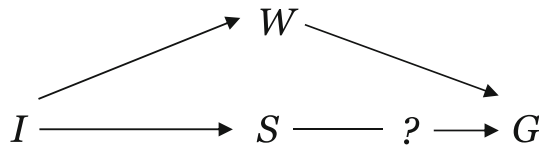


Fig. 3

Since there is a path from I to G in this model that does not go through S , it follows from (I3) that I is not an intervention variable for S with respect to G . Variables like W that are on paths to the purported effect variable that don't also go through the target variable are sometimes called 'confounders' in the experimental literature, and the need to ensure that our intervention variable is not also a cause of such variables is sometimes referred to as the need to 'control for confounders'.

(I4) is designed to address a similar problem. Suppose the money for the experiment was raised by cutting government funding for state schools. Let C be a variable such that $C = 1$ if the cuts occur and $C = 0$ otherwise. Since such cuts are likely to have an impact on the difference between private and state school grades, the causal model including this variable plausibly looks like this:

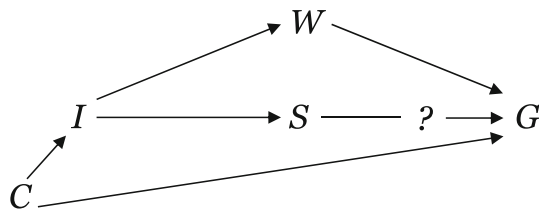


Fig. 4

(I4) rules out such scenarios. Since there is a path from C to G that doesn't go through S , (I4) requires that an intervention variable I for S with respect to G should

be statistically independent of C —and one way I can fail to be statistically independent of C is for C to be a cause of I .

There is one final problem with my fictional experiment: we can't be sure that the participants in the first group will actually *use* the money we give them to send their children to private schools. There might be other reasons—political objections to private education, for example—why these parents might decide to send their children to state schools after all, despite now having the resources to send them to private schools. (I2) is designed to address this problem. It's sometimes described as the requirement that an intervention be 'surgical', in Pearl's (2000) words. An intervention on X should amount to "lifting X from the influence of the old functional mechanism" in which it was embedded "and placing it under the influence of a new mechanism" (Pearl 2000, p. 70), one which "breaks whatever endogenous casual relationships are at work" (Woodward 2003, p. 135) in determining the value of X . In other words, an intervention on X should ensure that " X ceases to depend on the values of the other variables that cause X " (Woodward 2003, p. 98), so that its value "is determined completely by our intervention, the causal influence of the other variables being completely overridden" (Hitchcock 2012).¹²

If P is a variable representing the political commitments of the parents, the graph of the model containing I , P , S and G looks like this:

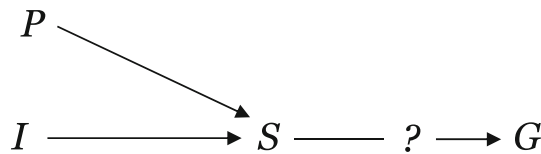


Fig. 5

According to (I2), I is an intervention variable for S with respect to G only if, when I takes its 'on' value, the value of I is the only actual cause of the value of S in this model. In particular, when $I = 1$, the value of P should *not* also be a cause of the value of S —changing the value of P should have no effect on the value of S . An intervention on whether a student attends private school, in other words, should ensure that the student attends private school *regardless* of the political commitments of her parents. Simply giving the parents a large sum of money fails to achieve this, because whether or not a child attends private school might continue to depend on the parents' political commitments even after the money has been handed over.

¹² Of course, in indeterministic models, it needn't follow from the fact that $I = 1$ is the only actual cause of the value of X that the value of X is 'completely determined' by the value of I . Since I am assuming a deterministic framework here for simplicity, however, I will occasionally allow myself the liberty of using these locutions interchangeably. Thanks to an anonymous referee for prompting me to clarify this.

4 Interventionism and Rational Causation

It's (*I2*)—the so-called ‘surgical constraint’—that Campbell thinks leads to trouble for causation in psychology. Suppose we're interested in whether S's intention to ϕ was a cause of her ϕ -ing. According to (AC), this is the case only if, had an intervention on whether or not S intends to ϕ been performed, there would have been a change in whether or not S ϕ -s. But what is it to intervene on whether someone has the intention to ϕ ?

We would naturally think of this in terms of providing someone with reasons to [ϕ], or reasons not to [ϕ]... The trouble with this is that it leaves intact the factors that are the usual causes of the someone's forming, or not forming, the intention to do something... This means that the intervention is not, in Pearl's term, ‘surgical’ (Campbell 2007, p. 61–2).

Simply telling S that ϕ -ing will make her happy doesn't count as an intervention on whether S intends to ϕ , for example, because it fails to satisfy (*I2*)—whether S intends to ϕ continues to depend on whether S believes that ϕ -ing will make her happy, even after the intervention is performed. So what, then, would an intervention on whether S intends to ϕ actually look like?

[Such an] intervention would have to come from outside and seize control of whether the subject had the intention, suspending the influence of the subject's usual reasons for forming an intention... This is evidently quite an unusual situation. (Campbell 2007, p. 62).

There are some philosophers who would deny that such an intervention on intentions is even possible. They would deny, in other words, that it's even coherent to speak of intentions to ϕ being manipulated independently of preferences or beliefs about the consequences of ϕ -ing. 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. As Campbell (2010, p. 71) characterizes the view, “[t]he mind has to be organized in a broadly rational way, for there to be a mind there at all”. There is no such thing, nor could there be such a thing, as an agent who believes that ϕ -ing will make her happy, has no reason not to ϕ , yet nevertheless fails to intend to ϕ , because it is an *a priori* prerequisite on something exemplifying mental states at all that their mental states fit together in a broadly rational way. In a slogan: one cannot simply pick-and-mix mental states. Therefore, a manipulation of whether or not an agent intends to ϕ which satisfies (*I2*) is conceptually impossible; and if *that's* right it follows from (AC) that intentions have no causal effects.

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 2007, p. 63). But Campbell himself is no friend of mental holism (see Campbell 2010). He grants that a surgical intervention on an intention to ϕ is *coherent*. His point is rather to emphasise just how strange such a thing would have

to be. What we are asking for is a manipulation of an agent's intentions that disrupts the sensitivity of her intentions to rational evaluation. Indeed:

Someone 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, p. 62).

Campbell concludes that “it is not credible that our interest in psychological causation is an interest in what would happen under such idealized conditions of alien control” (Campbell 2007, p. 62). Although he is happy to grant that the interventionist account is *extensionally adequate* (in that it doesn't misidentify a cause as a non-cause or vice versa), Campbell nevertheless considers it implausible to suppose that when we talk, say, of Jane's intention to dance as causing her dancing, we're saying something about what would have happened in cases so far removed from our own psychological lives as to be virtually unrecognisable.

In Sect. 6, I will describe what I think is the right solution to Campbell's problem. But first, I want to get clearer on what exactly the problem is. In particular, I want to argue that—contrary to what Campbell seems to suggest—the problem is not confined to causation in psychology.

5 The Problem of Abrupt Transitions

Suppose Tushar is driving in the outside lane of a two-lane road and realizes too late, at time t , that he needs to take the next exit.¹³ He misses the exit, and as such he is late for his meeting. Tushar's being in the outside lane at t was a cause of his lateness. According to (AC), this is so because had an intervention on which lane he was in at t been performed, there would have been a change in whether or not he was late for his meeting.

So what would it be to intervene on which lane Tushar was in at t ? One natural thought is something like this: we call him up some time before t and remind him that he needs to take the next exit. But it seems this manipulation wouldn't satisfy (I2). Whether Tushar was in the outside or the inside lane at $t-1$, after all, is a cause of whether he was in the outside or the inside lane at t in some model, for any sufficiently fine-grained units of time. An intervention on which lane Tushar was in at t should therefore remove this variable from under the influence of his location at any earlier time. In other words, the intervention must ensure that Tushar ends up in the inside lane at t , *wherever* he happens to be at any arbitrarily small time before that. One can certainly *imagine* interventions like that—we could place some kind of portal in the outside lane, for example, which has the effect of instantly

¹³ This case is taken from Woodward (2003, pp. 142–4), but was apparently originally discussed by David Lewis in an unpublished lecture.

teleporting Tushar into the inside lane if he happens 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 Tushar is late for his meeting, in violation of (I3)—for example, we'd probably have to wipe the memories of the drivers around Tushar, so that they don't become too alarmed at the sudden disappearance and reappearance of his car and crash into him, thereby making him late for his meeting.) But this would be a strange intervention indeed. Even if it's *coherent* to imagine a manipulation of Tushar's position at t which removes it from under the influence of his position at earlier times, and even if the interventionist account gets the right result in this case (that Tushar's being in the outside lane at t was a cause of his lateness), it still seems odd that when I talk of Tushar's being in the outside lane at t as a cause of his 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.

The case above involves variables making explicit reference to times. But the same problem arises in other models too. An intervention on whether or not a light turns on, for example, should ensure that it does so regardless of whether anyone flips the light switch, according to (I2); an 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 indeed the soil itself); and so on. If (I2) is applied at a sufficiently fine-grained level of detail, surgical interventions start to look *really* weird. Call this the *problem of abrupt transitions*. I submit that the problem Campbell identifies 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.¹⁴

It's notable that Lewis himself seemed to feel the force of the problem of abrupt transitions. When constructing his closeness metric on possible worlds, Lewis argued 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, p. 463).¹⁵ To determine whether Tushar's being in the outside lane at t caused him to be late for his meeting, for example, we should consider a possible world which starts diverging from the actual world a few seconds before t , and then smoothly transitions into a world in which he is in the inside lane at t . The interventionist account, however, seems inconsistent with such ‘transition periods’ between actual past and counterfactual future. Indeed, Woodward explicitly acknowledges that, “in contrast to Lewis, the interventionist account tells us that we should avoid transition periods entirely” (Woodward 2003, p. 144).

¹⁴ It may be that there is some additional strangeness involved in the idea of surgical interventions on mental states, of course, even if we reject mental holism. In saying that Campbell's problem is an instance of the problem of abrupt transitions, I mean only to suggest that both problems can be solved in the same way. Thanks to an anonymous referee for encouraging me to clarify this point.

¹⁵ Lewis's account “requires the thought of a run-up to the antecedent, a *ramp* from the actual world to the antecedent of the conditional”, to use Bennett's (2003, p. 214) characteristically evocative metaphor.

6 Indirect Interventions

Campbell's proposed solution to the problem of abrupt transitions is to abandon the surgical constraint on interventions altogether.¹⁶ But this, I want to suggest, is an overreaction. Recall why (I2) was introduced: it was designed to ensure that the value of the target variable can be fully determined by the value of the intervention variable. One of the reasons why increasing parental income is not an acceptable intervention on whether or not a child attends private school, for example, is the fact that the parents might have reasons for not sending their children to private school, such as political objections to private education, that have nothing to do with lack of funds. Here is the causal structure of this case again:

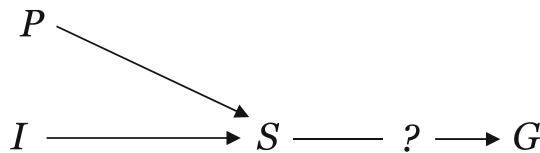


Fig. 5

Although I is a variable-level cause of S in this model, a change in the actual value of I may fail to result in a change in the actual value of S , due to interference from P .

(I2) addresses this issue by stipulating that for I to be an intervention variable for S , S must cease to depend on the values of *every* variable except I when $I = 1$. But this, in fact, is stronger than it needs to be. To illustrate, consider the example Campbell discusses of an ordinary manipulation of an intention. Let $N = 1$ if Jane intends to dance and 0 otherwise, $D = 1$ if Jane dances and 0 otherwise, $B = 1$ if Jane believes that dancing will make her happy and 0 otherwise, and $I = 1$ if I tell Jane that dancing will make her happy and 0 otherwise. The causal model containing these four variables looks like this:

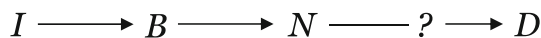


Fig. 6

Campbell correctly points out that I is not an intervention variable for N according to (I2), because when $I = 1$, the value of N continues to depend on the value of a variable besides I , namely B . But notice the difference between Figs. 5 and 6—in Fig. 5, P is part of a *separate* causal path, which threatens to disrupt or pre-empt the effect of the candidate intervention variable on the target variable. In Fig. 6, by contrast, B is *part* of the very causal path along which the

¹⁶ “[W]e can resolve this problem within a broadly interventionist framework, but... to do so we have to rethink our conception of an intervention; we have to move away from the focus on surgical interventions” (Campbell 2007, p. 63). Manipulations of variables satisfying (I1), (I3) and (I4), but not (I2), are sometimes called ‘soft’ interventions, and there is a burgeoning literature on how they can be used to learn about causal structures under certain assumptions; see Eberhardt and Scheines (2007), for example.

candidate intervention operates. That the value of N continues to depend on the value of B in Fig. 6 is no surprise—it *has* to, otherwise $I = 1$ wouldn't be an actual cause of the value of N in this model. Thus although the variable I should not count as an intervention variable for S in Fig. 5, it's not clear why I shouldn't be allowed to count as an intervention variable for N in Fig. 6.

This suggests the following revised definition of intervention variables:

INTERVENTION VARIABLE* (IV*): I is an intervention variable for X with respect to Y if and only if, in every model \mathbf{M} containing I , X and Y :

- (I1) I is a variable-level cause of X in \mathbf{M} ;
- (I2*) When I takes its 'on' value, $I = 1$ is an actual cause of the value of X in \mathbf{M} along some path P , and for every V in \mathbf{V} not on P , the value of V is not an actual cause of the value of X in \mathbf{M} ;
- (I3) Every path from I to Y in \mathbf{M} (if there is one at all) goes through X ;
- (I4) I is statistically independent of every variable in \mathbf{M} that is on a path to Y that does not go through X .

Both (I2) and (I2*) imply that I is not an intervention variable for S in Fig. 5. But although (I2) implies that I is not an intervention variable for N in Fig. 6 either, it's perfectly consistent with (I2*) that I is an intervention variable for N in Fig. 6, because even if the value of B is an actual cause of the value of N when I takes its 'on' value, B is on the path from I to N along which $I = 1$ is an actual cause of the value of N . Let's call a manipulation of X satisfying (I1), (I3) and (I4) a *direct* intervention if it satisfies (I2), and an *indirect* intervention if it satisfies (I2*) but not (I2).¹⁷ My claim is just that indirect interventions should be recognised as genuine interventions—the problem of abrupt transitions is a consequence of the fact that (I2) unnecessarily rules them out.

In defence of this claim, it's worth pointing out that very few (if any) manipulations actually carried out by practicing scientists—including manipulations of intentions—count as direct interventions. Consider, for instance, the studies reviewed by Webb and Sheeran's (2006) meta-analysis of the experimental evidence that intentions are causes of behaviour. Webb and Sheeran analysed forty-seven 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 0.36 on behaviour, showing, the authors conclude, that the correlation between intentions and behaviour is indeed causal. None of the 'interventions' in these studies satisfy (I2), however. 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), before urging the listener to carry out the procedure once a month. This was found to have

¹⁷ Note that the conditions on indirect interventions are still stronger than those on soft interventions – whereas under a soft intervention, the target variable may continue to depend on any of its variable-level causes, an indirect intervention must 'break' the connections between the target variable and all its causes *except* those on the path along which $I = 1$ is an actual cause of the value of the target variable.

an effect on both intentions and behaviour. But Brubaker and Fowler explicitly acknowledge that the ‘intervention’ affected the intentions of subjects *by* affecting the beliefs 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, p. 1414). Hence the ‘intervention’ failed to remove the target variable from under the influence of all its other causes, as (I2) requires. Needless to say, this apparent deficiency is not even remarked upon by the authors of the meta-analysis.

Of course, we shouldn’t necessarily expect the interventionist account to analyse causation in terms of the kinds of manipulations 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 of idealization in an account of the truth-conditions of causal claims in terms of the experimental procedures used to test them.¹⁸ But the real problem with (I2) is that it represents an ideal that scientists don’t seem to feel any pressure to even *try* and approach. Typically, an experiment in the special sciences 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 can be uniquely determined by the value of the intervention variable. They *won’t* typically be concerned to ensure that the intervention variable doesn’t act on the target variable ‘through’ any other variable—*except*, of course, if there is reason to think that that intermediate variable is a confounder.

This latter concern sometimes seems to be what Campbell and Woodward have in mind when they motivate (I2):

[S]uppose we leave intact the belief that [ϕ -ing] will make one happy. Then it is possible that the belief that [ϕ -ing] will make one happy causes both formation of the intention to [ϕ] and also directly causes performance of the action itself. In that case the intention to [ϕ] will be correlated with [ϕ -ing] even though the intention plays no role in causing the action. (Campbell 2007, pp. 61–2).

Similarly, Woodward insists on eliminating ‘transition periods’ between interventions and changes in the target variable, “because they may introduce factors that affect the effect independently of the putative cause” (Woodward 2003, pp. 144–5).

But this possibility is *already* taken care of by (I3). For example, if there is a path from whether Jane believes that dancing will make her happy (B) to whether she dances (D) that doesn’t go through whether she intends to dance (N) in some model containing these three variables, then (I3) already implies that an intervention variable for N with respect to D shouldn’t also be a cause of B in those models. If there is no such path, however, then there is no reason why an intervention on N with respect to D shouldn’t also be a cause of B . In other words, there seems to be

¹⁸ Again, the analogy with frequentism is helpful here – any plausible frequentist account should analyse objective probability in terms of *long-run* frequencies; longer, indeed, than anyone is in a practical position to determine.

no reason why an intervention cannot act ‘through’ other variables, so long as those variables are not independent causes of the purported effect variable.

Another purported motivation for (I2), which Woodward mentions in passing, has to do with backward causation. According to Lewis (1979), the closest possible world in which Tushar isn’t in the outside lane at t is one in which a small ‘miracle’ occurs in his brain a few seconds before t , after which he smoothly transitions into the inside lane. But this seems to imply that, if Tushar hadn’t been in the outside lane at t , a miracle would have occurred a few seconds before t —and it follows from *this*, Woodward argues, that on Lewis’s counterfactual account of causation, “we get backward causation in a case in which backward causation is clearly not at work” (Woodward 2003, p. 143).¹⁹ Woodward seems to think that (I2), by eliminating ‘transition periods’, also deals with this problem of widespread backwards causation.

But again, this issue is *already* taken care of by (I3). Consider Fig. 6 again:

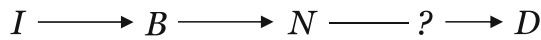


Fig. 6

It’s consistent with (I2*) that I is an intervention variable for N with respect to D . But it’s not consistent with (I3) that I is an intervention variable for N with respect to B , because there is a path in this model from I to B which does not go through N . Thus it doesn’t follow from the fact that a change in I results in a change in B that N is a cause of B —in other words, it doesn’t follow from the existence of a ‘transition period’ between an intervention and a change in the target variable that the change in the target variable is a cause of the earlier events in the transition period, contrary to what Woodward seems to suggest.

In summary, I think there are no good reasons to rule out indirect interventions in our definitions of causation. They are consistent with the original motivations for introducing the surgical constraint, they are frequently performed in science, and they don’t lead to widespread backward causation. The problem of abrupt transitions is a consequence of the fact that (I2) is stronger than it needs to be.

7 Rational Causation Revisited

So far I have argued that (I2*), and not (I2), is the correct constraint on interventions to adopt. This means that a manipulation of a subject’s intentions which proceeds by way of telling her what will or will not make her happy may well count as an intervention, even though her intentions will continue to depend on their usual rational causes. And *this* means that our interest in the causal consequences of

¹⁹ Lewis has a response to this objection, which Woodward doesn’t acknowledge: “There may be a variety of ways the transition might go, hence there may be no true counterfactuals that say in any detail how the immediate past would be if the present were different. I hope not, since if there were a definite and detailed dependence, it would be hard for me to say why some of this dependence should not be interpreted—wrongly, of course—as backward causation over short intervals of time in cases that are not at all extraordinary.” (Lewis 1979, p. 463). For discussion of this response, see Dorr (2016, pp. 262–3).

mental states is not necessarily an interest in the behaviour of agents lacking rational autonomy.

There are exceptions to this general rule, however, even after we replace (I2) with (I2*). Suppose that Johnny is a secret agent, who forms a belief that there is poison in his wine.²⁰ This causes him to form an intention to drop (and therefore smash) his wineglass, 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 wine is poisoned to whether or not he drops the glass, which doesn't go through whether or not Johnny intends to drop his glass. An intervention on whether Johnny intends to drop his glass with respect to whether he drops his glass must therefore, given (I3), manipulate his intention without affecting whether or not he believes the wine is poisoned; that is, without affecting the rational cause of Johnny's intention. So in *some* cases, interventions involving the partial suspension of an agent's rational autonomy may be required to establish a causal connection between intentions and behaviour, albeit because of (I3) and not (I2*). But these are special cases. Indeed, one could argue that cases like these strike us as odd or problematic precisely *because* of the weirdness of the interventions required to establish the causal claims in question. The point is that there is an important distinction between these kinds of cases and everyday cases of psychological causation, one over which the surgical constraint's blanket ban on indirect interventions runs roughshod, but which can be captured by my revised set of conditions on interventions.

8 Conclusion

In this paper, I've argued that the problem Campbell raises for interventionist approaches to causation in psychology is an instance of a wider problem, the problem of abrupt transitions, which arises from the so-called 'surgical constraint' on interventions. The correct solution to this problem is not to abandon the surgical constraint entirely, as Campbell recommends, but rather to replace it with a weaker constraint, one which is consistent with the possibility of indirect interventions, but which nevertheless does all the work the surgical constraint was designed to do. On this revised version of interventionism, it doesn't follow—except perhaps in certain interesting special cases—that our interest in psychological causation is an interest in the behaviour of agents lacking rational autonomy.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

²⁰ For classic examples of these kinds of cases, see Chisholm (1966) and Davidson (2001, p. 79).

References

- Baumgartner, M. (2013). Rendering interventionism and non-reductive physicalism compatible. *Dialectica*, 67, 1–27.
- Baumgartner, M. (2017). The inherent empirical underdetermination of mental causation. *Australasian Journal of Philosophy*, 96, 335–350.
- Bennett, J. (2003). *A philosophical guide to conditionals*. Oxford: Oxford University Press.
- Brubaker, R. G., & Fowler, C. (1990). Encouraging college males to perform testicular self-examination: Evaluation of a persuasive message based on the revised theory of reasoned action. *Journal of Applied Social Psychology*, 20, 1411–1422.
- Campbell, J. (2007). An interventionist approach to causation in psychology. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (pp. 58–66). Oxford: Oxford University Press.
- Campbell, J. (2010). Independence of variables in mental causation. *Philosophical Issues*, 20, 64–79.
- Chisholm, R. (1966). Freedom and action. In K. Lehrer (Ed.), *Freedom and determinism* (pp. 11–44). New York: Random House.
- Coleman, J., & Hoffer, T. (1987). *Public and private high schools*. New York: Basic Books.
- Davidson, D. (2001). Freedom to act. In D. Davidson (Ed.), *Essays on actions and events* (2nd ed., pp. 63–81). Oxford: Oxford University Press.
- Dorr, C. (2016). Against counterfactual miracles. *Philosophical Review*, 125, 241–286.
- Eberhardt, F., & Scheines, R. (2007). Interventions and causal inference. *Philosophy of Science*, 74, 981–995.
- Franklin-Hall, L. (2016). High-level explanation and the interventionist's variables problem. *British Journal for the Philosophy of Science*, 67, 553–577.
- Halpern, J. Y., & Pearl, J. (2005). Causes and explanations: A structural-model approach. Part I—causes. *British Journal for the Philosophy of Science*, 56, 843–887.
- Hesslow, G. (1976). Two notes on the probabilistic approach to causality. *Philosophy of Science*, 43, 290–292.
- Hitchcock, C. (2001). The intransitivity of causation revealed in equations and graphs. *The Journal of Philosophy*, 98, 273–299.
- Hitchcock, C. (2012). Probabilistic causation. In E. N. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Winter 2012 Edition). <https://plato.stanford.edu/archives/win2012/entries/causation-probabilistic/>.
- Jackson, F. (1998). *From metaphysics to ethics: A defence of conceptual analysis*. Oxford: Oxford University Press.
- Leslie, S. (2012). Generics. In G. Russell & D. Fara (Eds.), *The Routledge handbook to philosophy of language* (pp. 355–366). London: Routledge.
- Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy*, 67, 427–446.
- Lewis, D. (1979). Counterfactual dependence and time's arrow. *Noûs*, 13, 455–476.
- Menzies, P. (2008). The exclusion problem, the determination relation, and contrastive causation. In J. Hohwy & J. Kallestrup (Eds.), *Being reduced: New essays on reduction, explanation, and causation* (pp. 196–217). Oxford: Oxford University Press.
- Pearl, J. (2000). *Causality: Models, reasoning and inference*. New York: Cambridge University Press.
- Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search*. Cambridge, MA: The MIT Press.
- Wallace, D. (2012). *The emergent multiverse: Quantum theory according to the Everett interpretation*. Oxford: Oxford University Press.
- Webb, T. L., & Sheeran, P. (2006). Does changing behavioral intentions engender behavior change? A meta-analysis of the experimental evidence. *Psychological Bulletin*, 132, 249–268.
- Weslake, B. (forthcoming). Exclusion excluded. *International Studies in the Philosophy of Science*.
- Woodward, J. (2003). *Making things happen: A theory of causal explanation*. Oxford: Oxford University Press.
- Woodward, J. (2008). Response to Strevens. *Philosophy and Phenomenological Research*, 77, 193–212.