Synthese (2014) 191:1775–1791
DOI 10.1007/s11229-013-0366-1

How agency can solve interventionism's problem of circularity

Victor Gijsbers · Leon de Bruin

Received: 18 January 2012 / Accepted: 22 October 2012 / Published online: 21 November 2013

Abstract Woodward's interventionist theory of causation is beset by a problem of circularity: the analysis of causes is in terms of interventions, and the analysis of interventions is in terms of causes. This is not in itself an argument against the correctness of the analysis. But by requiring us to have causal knowledge prior to making any judgements about causation, Woodward's theory does make it mysterious how we can ever start acquiring causal knowledge. We present a solution to this problem by showing how the interventionist notion of causation can be rationally generated from a more primitive agency notion of causation. The agency notion is easily and non-circularly applicable, but fails when we attempt to capture causal relations between non-actions. We show that the interventionist notion of causation serves as an appropriate generalisation of the agency notion. Furthermore, the causal judgements based on the latter generally remain true when rephrased in terms of the former, which allows one to use the causal knowledge gained by applying the agency notion as a basis for applying Woodward's interventionist theory. We then present an overview of relevant empirical evidence from developmental psychology which shows that our proposed rational reconstruction lines up neatly with the actual development of causal reasoning in children. This gives additional plausibility to our proposal. The article thus provides a solution to one of the main problems of interventionism while keeping Woodward's analysis intact.

Keywords Causation · Intervention · Agency · Circularity · Woodward

V. Gijsbers (✉)
Instituut voor Wijsbegeerte, Universiteit Leiden, Postbus 9515, 2300 RA Leiden, The Netherlands
e-mail: V.Gijsbers@hum.leidenuniv.nl

L. de Bruin
Institut für Philosophie II, Ruhr-Universität Bochum, Universitätsstr. 150, 44801 Bochum, Germany
e-mail: lcdebruin@gmail.com

1 Introduction

The analysis of causation has been an important and controversial topic in the philosophy of science for many decades (see, for instance, Sosa and Tooley 1993; Psillos 2002; Beebee et al. 2009). An approach that has achieved wide popularity recently is James Woodward's interventionism (Woodward 2003a). The main idea of Woodward's theory is that causation should be analysed in terms of intervention: roughly, X causes Y if and only if there is a possible intervention I on X that changes the value of Y. Much of the theory's work is done by conditions which spell out exactly which interactions count as interventions.

If one looks at the intuitions underlying the theory, interventionism is closely related to earlier agency theories of causation (e. g., Collingwood 1940; Von Wright 1971; Menzies and Price 1993), which analysed causation in terms of human action. However, interventionism distinguishes itself from these theories by avoiding the concept of agency. This makes it immune to what has generally been seen as the main problem faced by agency theories, namely, that it is not clear how a theory that analyses causation in terms of agency can handle causal relations between events that humans could not possibly cause, such as earthquakes or, even more radically, the Big Bang (Menzies and Price 1993, Sect. 5; Woodward 2003b, pp. 123–127; 2008a, Sect. 3).

Unsurprisingly, Woodward's analysis of causation has some problems of its own. To our mind, the most fundamental criticism that has been levelled against it is that there is an infinite regress in the interventionist definition of cause: Woodward's definition of *cause* contains the term *intervention*, while his definition of *intervention* contains the term *cause*. Woodward recognises this circularity and argues that it is not vicious. However, several writers point out that even if it is not vicious in certain respects, it nevertheless raises tough problems for his theory (Glymour 2004; de Regt 2004; Baumgartner 2009).

In this article, we first wish to argue that the circularity in the interventionist theory is indeed problematic, but that it is not a problem of *analysis*, but a problem of *genesis*. That is, we will argue that although Woodward can hold that his theory captures the *meaning* of causation, the theory nevertheless makes it highly mysterious how we could ever *acquire* such a concept and start gathering causal knowledge. Since we appear to *have* causal knowledge, this is problematic. The existence of such a mystery casts doubt upon the analysis itself, especially since other theories of causation—for instance, theories that are based on observed correlations—do not have a problem of genesis.

The main aim of our article is to demonstrate that the problem of genesis can be solved within the framework of interventionism by assuming that our interventionist notion of causation develops from an agency notion of causation. We provide a rational reconstruction of this development. Furthermore, we show that, from an empirical point of view, it is plausible that something like our reconstruction takes place during the early life of every human being.

Defenders of interventionism can adopt this solution to the problem of circularity without having to change their analysis of causation. We claim that agency plays an important role in the development of the concept of causation, but that it has been superseded and generalised away in the concept of causation as we currently have

it. Our own actual sympathies for agency theories run deeper than that, but we will not press them in this article. Here, we present ourselves as Woodwardians using the notion of agency to solve a difficulty of the interventionist theory.

The argument is divided into three parts. In Sect. 2 we discuss the problem of circularity in Woodward's theory, and conclude that the real problem facing interventionism is a problem of genesis. At the same time, we develop a rough idea of how this problem can be solved without changing the interventionist analysis of causation itself. We then give a rational reconstruction of the genesis of the interventionist notion of causation from a simple agency theory in Sect. 3. Finally, in Sect. 4 we give an overview of the empirical evidence for this proposed story of genesis. All in all, we present what we believe to be a philosophically pleasing and empirically plausible solution to the main problem besetting Woodward's interventionism.

2 The problem of circularity

According to Woodward, "X is a contributing cause to Y with respect to variable set $\mathbf{V}$" means (roughly) that there is a possible intervention on X with respect to Y that changes the value of Y while certain other variables are held fixed (Woodward 2003a, p. 59). Interventions are defined in terms of intervention variables; and a necessary condition for I being an intervention variable for X with respect to Y is that I is a contributing cause of X (Woodward 2003b, p. 98). I is a contributing cause to X just in case there is some variable set $\mathbf{V}$ such that I is a contributing cause to X with respect to $\mathbf{V}$ (Woodward 2008b, p. 209). Thus, we have a circle of definitions moving from "contributing cause with respect to a variable set", to "intervention", to "intervention variable", to "contributing cause" to "contributing cause with respect to a variable set". Is this a problem, and if so, what kind of problem is it?

Here is one possible line of argument. The main philosophical problem with causation is the fact that it is or seems to be a modal notion, and modality is mysterious. So what we are looking for when we analyse causation is a reductive analysis in non-modal terms. Potential candidates are Humean and Mill–Ramsey–Lewis-style theories; but interventionism evidently fails to deliver the goods, because its circularity ensures that the modal notions are not analysed away. We thus have a *failure of reduction.*

While adherents of certain forms of positivism and empiricism may find this line of argument persuasive, it is clear that an interventionist need not be too worried about it. Woodward specifically states (2003b, p. 106) that he does not believe a reductive theory of causation to be possible, and cites the long history of failures of Humean and correlation-based theories as a reason for this belief. His theory is supposed to show us how different causal concepts hang together and how different causal claims are related; it is not supposed to show us how causation can be reduced to non-modal concepts (as the use of counterfactual conditions involving "possible" interventions already shows). Conceived in this way, then, the problem of circularity is simply irrelevant to interventionism; those who invoke it would be begging the question.

Here is another line of argument. A theory of causation should tell us what "X causes Y" means. It evidently cannot do this by simply repeating the phrase and stating that "X causes Y just in case X causes Y". But even if the analysis is more enlightening

than that, and does for instance start out by telling us that X causes Y if and only if there is a possible intervention I on X that changes Y, it would still be disappointing if it turns out that part of the analysis of "there is an intervention I on X that changes Y" is "X causes Y". Such a theory would suffer from a *circular regress*. It would not necessarily be worthless, as it would show us how the concepts of cause and of intervention hang together; but it could hardly be thought to be a satisfying analysis of the meaning of causal claims. It would also be epistemologically troubling, for we would seem to need to know whether or not X causes Y before we could know whether X causes Y.

Woodward recognises this problem, but is quick to point out that his theory does not suffer from such a circular regress (2003b, p. 105). Part of the analysis of "X causes Y" is that there is a possible intervention I on X with respect to Y that changes the value of Y; and part of the analysis of that claim is that "I causes X". Since the latter causal claim is not identical to the former, there is no circular regress. Speaking in terms of knowledge, although we need some causal knowledge before we can come to know a new causal fact, we do not need to know that very fact itself prior to coming to know it.

But this suggests a third line of argument. According to interventionism, it is part of the meaning of "X causes Y" that there is some possible I such that (among other things) I causes X. But part of the meaning of that latter claim is that there is some possible J such that (among other things) J causes I. And so on, *ad infinitum*, unless at some point we enter a circle. So supposing that Woodward is right and there is no circular regress in his theory, there is nevertheless an *infinite regress*. Is this problematic?

Woodward does not think it is. According to him, the infinite regress simply means that causation cannot be defined in non-causal terms, which is not a problem since he does not want to give a reductive analysis of causation anyway. The infinite regress also means that we cannot get causal knowledge and cannot perform causal tests without having some previous causal knowledge. But this too is a consequence that Woodward simply accepts: all methods of causal reasoning and testing either require prior causal knowledge or stipulate causal assumptions. No causes in, no causes out, is the slogan (as Cartwright formulated it in her 1994); and if that slogan holds true for all methods of causal reasoning, then interventionism is no worse off than any of its competitors.

These arguments have not been accepted by all commentators. For instance, Glymour and De Regt insist that the infinite regress still engenders what we will call a *problem of analysis*; that is, the infinite regress shows that interventionism cannot adequately capture the meaning of the term causation. De Regt (2004) writes:

> MT [Woodward's theory about the meaning of causal claims] does not reduce causation to other concepts, because causal relations are defined via the notion of intervention, which is itself a causal notion. Woodward argues that this circularity is not vicious because the causal information required to characterize the intervention is independent of the alleged causal relation between X and Y... However, this argument holds water only if MT is regarded as a theory of causal inference or testing. If MT is a theory of the meaning of causal claims, then it is hard to see how the circularity cannot be vicious.

Glymour (2004) makes essentially the same point:

> Ok, the definition is ill-founded, not circular: it could never be applied to determine direct causes ab initio. It could tell us something fundamental about how our notions of cause and intervention are often related, but it cannot be an analysis of the very meanings – whatever those are – of 'intervention' and 'direct cause'. (p. 785)

De Regt and Glymour both claim that an analysis of causation which leads to an infinite regress cannot be a satisfactory analysis of the meaning of the term, because such a definition is in some sense viciously circular and cannot determine causes ab initio. It is easy to understand the underlying intuition here. Suppose we analyse the notion of a natural number in the following way: "if n is a natural number, then the successor of n is also a natural number". This analysis is incomplete. It will not allow us to determine whether something is a natural number until we add a way of getting the iterative process going, for example, by adding that "0 is a natural number". Woodward has given us an analysis of the form "if m is a cause of n, and $X(m, n, o)$, then n is a cause of o". This seems to be incomplete unless we are given a way to start the iterative process.

The analogy with natural numbers is suggestive. But are regressing analyses necessarily incomplete? Let us consider, as something of a toy example, the idea that we can analyse "law" (in the legal sense) as "a rule stipulated to be a law by the legislature". Furthermore, we define the legislature as "the body of people who have lawfully been given the authority to make laws", and we define that authority is lawfully given just in case it is "given in accordance to the laws".

There is an obvious circle here: the legislature decides which rules are laws, and the laws designate which people are the legislature. Our analysis cannot be used to decide which things are laws ab initio. Does this mean that the analysis is incomplete, and that we must add a rule saying which were the first laws or what was the first legislature?

Perhaps we do. But it certainly seems possible to hold that our analysis exhausts the concept of law. It is neither inconsistent nor incoherent to claim that the first legislature turned itself into that first legislature by passing the first law. If there is a difficulty with such a proposal, it is not that it is absurd, but that it leaves open the question of how we can *know* that something is a law or a legislature. It seems that we must already know that some things are laws or legislatures before we can decide of anything else that it is a law or a legislature.

Does such an epistemic problem prove that our analysis of the meaning of law is wrong? Only if we assume that an analysis of the meaning of a term consists in giving a method to decide whether an object falls under that term or not. Such an assumption seems dangerously close to a verification theory of meaning, and Woodward is under no obligation to accept it. The interventionist, then, can flat out deny that the circularity in Woodward's definition leads to a problem of analysis.

But perhaps the problem of circularity should not be formulated in terms of meaning. Baumgartner (2009) claims that the infinite regress in interventionism leads to a problem with testing causal claims, since you always need prior causal knowledge to

perform a test. If so, how could such testing ever get started? For Baumgartner, the problem with interventionism's circularity is not primarily a problem of analysis; it is a *problem of genesis* (this is our term, not his). Given interventionism, it is unclear how we could ever get any causal knowledge, how our causal discourse could ever *get started*. It appears that we always already need to have some causal knowledge before we could gain any.

Baumgartner describes two strategies which Woodward could use to solve this problem. The first is to claim that we do have prior causal knowledge; the second is to simply stipulate some variables as causes of (or intervention variables for) others. Both strategies, Baumgartner argues, are problematic. The latter, where we simply stipulate that a certain set of causal claims is true, works well if you are only interested in the *validity* of causal inferences, but using it means giving up all hope of showing which causal inferences are *sound* (i.e., not only valid but also true). An interventionist who would be content with having a theory of causal inference could adopt this strategy, but an interventionist who wants to be able to explain how we can get any true (or even simply justified) causal beliefs in the first place cannot.

So if Woodward wants to stop the infinite regress of testing, and explain how causal discourse can get started, he must claim that we can have causal knowledge prior to applying the interventionist criterion. It is possible to have prior causal knowledge, Baumgartner continues, if we have some suitable, independent heuristic that allows us to decide for some A and B that A causes B. There should, in other words, be some property of pairs A and B that, while not being a part of the interventionist analysis of causation, nevertheless is strongly correlated with A causing B. We could, for instance, adopt the heuristic "you are justified in concluding that A causes B if B happened near and shortly after A", and then use the beliefs gained this way as a basis for applying the interventionist theory. (Of course, this particular example does not work, since it leads to wrong conclusions far too often.)

Baumgartner allows that there might be a good heuristic to be found. But he casts doubt on the idea that a non-interventionist condition could be interpreted as *merely* a heuristic.

> [I]n order to establish a certain non-definitional property of an entity of type t as a heuristic measure for the identification of entities of type t, it must be shown that the non-definitional property indeed *coincides* with the definitional properties of entities of type t. That is only possible if at least some entities of type t can actually be identified explicitly by applying t's definition. That is, heuristics for t can only be validated if the definition of entities of type t is applicable in a finite number of steps, at least in principle. (p. 186; emphasis in the original.)

But the regress problem arises precisely because this last condition is not met by Woodward's analysis. Thus, Baumgartner concludes:

> In view of this lack of a single positive application of [Woodward's analysis], non-interventionist accounts cannot be given the status of heuristics for assessing the satisfaction of [Woodward's analysis]. Instead, they provide *self-contained*

> *analyses* of causation that are *independent* of the notion of intervention. (p. 186; emphasis in the original.)

Baumgartner's point is that because Woodward's theory cannot be applied in a finite number of steps to even one single variable set, an interventionist can never check whether the claims of any other account of causation coincide with the claims of the interventionist theory. This means that the interventionist cannot justify using any particular account of causation as a heuristic, except by accepting it as a self-contained, independent analysis of causation.

But if one does so, what purpose is the interventionist theory supposed to serve? Well, it might be the case that the non-interventionist account of causation, while independent of interventionism, is to some extent unsatisfactory, and that interventionism is introduced to eliminate these deficiencies. The body of causal claims generated by the original account could then be used as the basis of the interventionist theory, although it would be a basis that had to be extended and/or corrected by intervenionism. For instance, suppose that we start out with some probabilistic account of causation. It might turn out that the causal claims we accept based on that theory can be more usefully generalised and systematised if we use an interventionist theory. By using the results of the probabilistic account as at least prima facie justified, it might be possible to bootstrap ourselves from the probabilistic account to the interventionist one. In this case, the probabilistic account does not function as a *heuristic* for the interventionist theory, but as a (temporal and, more importantly, methodological) *precursor* to that theory.

However, if the only way to get started with the interventionist theory of causation is to first accept a theory of causation that is conceptually wholly independent of the interventionist theory, causation becomes a mysteriously dual notion. If we need to start our causal discourse by understanding causation in terms of probabilities, it would then be puzzling if causation turned out to be an irreducibly interventionist notion. This mysterious duality can only be avoided if the original theory is not just a methodological but also a conceptual precursor of interventionism; that is, if we can understand interventionism as a development and refinement of the original theory.

These points are crucial to an understanding of our project, so let us phrase them once again in a slightly different way. Baumgartner claims (correctly, in our view) that the only way to validate A as a heuristic for B is by comparing decisions based on A and decisions based on B, and seeing that they coincide. But if we have developed B as a theory that allows us to generalise and gain better understanding of the results first gained by applying A, there is no need to validate A as a heuristic of B. Rather, B has to be validated as a better, more useful analysis of the notion that was first introduced by A. The genesis of the interventionist notion of causation can then be thought of in the following way. First, we start out with a simple, easily applicable notion of causation, A. With A in hand, we start gathering causal knowledge. But for one reason or another—which would of course have to be spelled out—A turns out to be unsatisfactory. We recognise that we could adopt a causal notion B that allows us to keep almost all our previous causal knowledge and to extend it in useful ways, while the conceptual continuity between A and B still allows us to think of B as a theory of

causation. We thus come to adopt B, using our previous causal knowledge—perhaps with some objectionable elements purged from it—as an inference basis for future causal testing.[1]

In the rest of this paper, we wish to defend the view that this scenario comes close to what actually happens. The interventionist notion of causation can be understood as a sophisticated version of a primitive agency notion of causation, a version we have arrived at after generalising away from human action. The interventionist theory can escape the infinite regress of testing because it can treat the original causal claims of the agency theory as prima facie justified claims about contributing causation simpliciter—they may turn out to be wrong on occasion, but they are in fact (that is, contingently) mostly right. The relation between interventionism and agency is not necessary, but turns out to be strong enough in our actual world to allow for bootstrapping from the latter to the former. In the next section, we will show how this bootstrapping takes place.

3 From agency to intervention

How can an interventionist notion of causation be generated from an agency notion of causation? In the current section, we will give a rational reconstruction of this process, presenting first a rough outline and then describing the individual steps in more detail. For the sake of brevity, we will abbreviate the agency notion of causation as causation$_1$, and the interventionist notion as causation$_2$. Causation$_2$ is Woodward's notion of a contributing cause simpliciter (that is, not relativised to any variable set).

Our rational reconstruction consists of the defence of three claims:

(1) A simple agency theory can be applied ab initio to gain justified causal$_1$ beliefs. (That is: the concept causation$_1$ can be applied without presupposing prior causal knowledge. If that were not the case, we would just have replaced one problem of genesis with another.)
(2) Causation$_1$ is an unsatisfying concept, because there are some obvious and useful ways to generalise our causal$_1$ knowledge that are impossible to capture in terms of agency. Causation$_2$ can be understood as a generalisation of causation$_1$ that disposes with the concept of agency.
(3) Most claims that are true when formulated in terms of causation$_1$ can be and are in fact held true when formulated in terms of causation$_2$.

Claim 3 needs a little more explication. We have seen that the interventionist theory, which defines causation$_2$, can only be applied if we start with a prior body of causal knowledge that can then be extended and perhaps partly revised. In our reconstruction of the genesis of causation$_2$, this body of knowledge is the body of causal$_1$ knowledge. If this reconstruction is correct, two things follow. First, the body of causal$_1$ knowledge must fit (with at most minor exceptions) into the formal structure of Woodward's

[1] Throughout this paper, we will assume that analyses of the concept of causation can be ranked as better and worse on pragmatic grounds, without the need for a theory-neutral set of intuitions underlying them. We also assume that it is possible to see intellectual progress from a worse to a better theory from a purely internal perspective, without the need for a further point of view that is neutral with respect to the two theories. A defence of these assumptions falls outside the scope of the current paper.

theory. While this theory underdetermines for each particular causal claim whether it is true or false, it does determine for some sets of causal claims that not all of its members can be true; or that if all of its members are true, some other particular causal statement must also be true. A body of purported causal knowledge may or may not fit these constraints. Claim 3 states that the body of causal$_1$ knowledge does fit interventionism's constraints; that is, most causal$_1$ knowledge can be held true when formulated in terms of causation$_2$.

Second, claim 3 also states that most of this body is in fact held true. That is, most of the claims made by the agency theory are causal claims that we actually accept (and that the interventionist thus wishes to capture with the notion of causation$_2$). If this were not the case, we apparently did not use causation$_1$ as a basis for the concept of causation that interventionism is meant to explicate.

According to our reconstruction, then, there are two kinds of continuity between causation$_1$ and causation$_2$. First, there is a conceptual continuity: the interventionist theory can be seen as a rational development of the agency theory. Second, there is an extensional continuity: most claims that come out true under the agency theory can also be held true by those who use an interventionist theory, and are in fact held true by us. As explained in Sect. 2, the first kind of continuity is needed to circumvent Baumgartner's criticism by establishing that the agency theory is not a heuristic for, but a *conceptual precursor* of the interventionist theory. The second kind of continuity, meanwhile, is needed to establish that the agency theory can function as a *methodological precursor* to the interventionist theory.

Our reconstruction will of course be sketchy. It is unlikely that the actual development of our notion of causation can be neatly divided into an agency stage and an interventionist stage, or proceeds based on the rational arguments that we will present. Also, we will not claim that Woodward's interventionism is the only possible development of an agency theory, or that the agency theory is the only possible precursor for interventionism. Rather, the aim of the reconstruction is to show that it is possible, in spite of the theory's circularity, to have justified causal$_2$ beliefs. In Sect. 4 we will say more about how the development of our causal notions actually takes place, and we will see that it roughly lines up with our proposed reconstruction.

Let us start by giving a more precise definition of causation$_1$. A simple version of the agency theory of causation can be found in the work of Von Wright:

> p is a cause relative to q, and q an effect relative to p, if and only if by doing p we could bring about q or by suppressing p we could remove q or prevent it from happening. (Von Wright 1971, p. 70)

In our discussion, we will assume that the variables (which we will call X and Y for consistency with Woodward's use) stand for type-level events, except where context makes it clear that we speak of particular instantiations of those types. Moreover, the first type-level event (the cause) is *always* a type of action. This makes the theory quite restricted in scope: it does not make claims about causal relations between non-actions.

We will understand type-level causal claims as statistical claims in the sense that they do not have to be exceptionless. Finally, we will assume that "doing" (or "taking an action") is a primitive notion. In particular, that "S does X" is not to be interpreted

as the claim that there is a causal connection between two events, say a mental state of S and an event X. It is also not to be understood as the claim that S "intervenes on" X in the Woodwardian sense.

It is easy to see why this notion of causation$_1$ would be useful to agents. We desire certain events to happen or not to happen. If we can perform some action X that makes it likelier that a desirable event Y happens, we would like to know about this connection. And if action X would prevent an undesirable event Y, that too would be good to know. In fact, it is hard to think of a concept that is more useful for agents to have than causation$_1$.

Let us move on to the defense of the first claim. Is it true that the simple agency theory can be applied ab initio to gain justified causal$_1$ beliefs?

One worry can be put aside easily, namely the worry that "bring about" and "prevent" are causal notions, and that Von Wright's definition is therefore circular. We already pointed out that we use these terms as primitives, not as notions that are somehow defined in terms of causation. Furthermore, we already saw, in Sect. 2, that the aim of our theory is not the reduction of a modal notion of causation to non-modal notions, so the appearance of modal concepts in Von Wright's definition of causation cannot count against it.

But this is merely a negative point. In order to defend our first claim, we must provide an epistemic procedure that allows us to gain justified beliefs about causal$_1$ claims. The agent begins by noticing a correlation between the action X and the event Y: when the agent does X, Y tends to happen more often than when the agent does not do X. When I clap my hands, I tend to hear a loud noise. When I kick a tree full of apples, apples tend to fall out. When I yawn, the sun tends to set.

Such a correlation of course does not prove that I can bring about Y by doing X. But there are easy ways to test this hypothesis. I can do X in circumstances in which I otherwise would not do X, and see if this does or does not break the correlation with Y. I can refrain from doing X in circumstances in which I normally do it, and see whether Y still happens. Normally, I kick trees only in the morning, when I am hungry. This time, I decide to kick trees at random throughout the day, and it turns out that apples still fall out whenever I kick. Normally, I start yawning when I have been awake for many hours; this time I refrain from yawning, but the sun still sets. These tests give me some justification for believing that by kicking, I bring about the falling of fruit; and that by yawning, I do not bring about the setting of the sun.

Why are we allowed to speak of justification? If there is a causal$_1$ relation between X and Y, then I expect the correlation between the two variables to remain when I change my X-behaviour; for such a correlation is implied by the claim that doing X brings about Y. On the other hand, if there is no causal$_1$ relation between X and Y, we do not expect the correlation to remain under variations of my X-behaviour. (Though it would perhaps be too strong to claim that we expect the correlation to disappear.) So if I do change my X-behaviour and the correlation remains, this raises the Bayesian probability of there being a causal$_1$ relation between X and Y.[2]

[2] We are not committing ourselves to Bayesianism here: other theories of confirmation would yield the same verdict.

Can we really expect people who do not even have the concept of causation to engage in such testing procedures? Perhaps not consciously, not if that implies that they have a theory of testing and justification. But something like the procedure described is what happens naturally to a reasoning agent whose whims and sudden inclinations ensure that his pattern of actions is not rigid. The procedure we have suggested is therefore not psychologically implausible. (See also Woodward 2007, and Sect. 4 of this article.)

Agents can thus gain a body of justified causal$_1$ beliefs. But why would they bother to develop their notion of causation towards the interventionist idea of causation$_2$? And are most of the justified causal$_1$ beliefs also held to be true causal$_2$ beliefs, which is necessary for our bootstrap procedure to work?

Extending the notion of causation$_1$ to encompass non-actions as causes will be very useful, since complex causal reasoning about actions requires such an extension. As a simple example, we would like to be able to reason as follows: action X causes$_1$ non-action event Y; Y causes$_1$ non-action event Z; so X causes$_1$ Z, that is, by performing X I can make Z happen. But this kind of reasoning is impossible on the agency theory (as we have presented it), because only actions can appear on the left hand side of causation$_1$. This is a serious limitation of the agency theory.

Knowledge about the causal$_1$ relations between actions and events can therefore be systematised and made more efficient when the notion of causation is extended to relations between non-action events.[3] A generalising move away from agency is therefore to be expected. But why would this move be towards an interventionist concept of causation?

Before we answer that question, it is useful to look at the truth of the justified causation$_1$ beliefs when they are transposed to an interventionist causation$_2$ vocabulary. The epistemic procedure we described for forming justified causal$_1$ beliefs was based on two things: correlations and non-rigid patterns of behaviour. Why would this procedure uncover what we see as true causal$_2$ knowledge, rather than falling into the pitfalls of spurious correlation that appear whenever we have common cause structures?

The answer is that in fact (contingently) our actions, and especially the more experimental ones that break with ingrained patterns of behaviour, are almost never related to their seeming results through a common cause structure. It is almost never the case that a class of events A causes both my action X and its seeming result Y. Not because such cases are impossible; they are in fact possible. As an example, assume that there is a correlation between a state B of your brain and you raising your hand five seconds later; and assume furthermore that B causes a beep to be heard six seconds later through some causal path that does not go through the raising of your hand.Such a scenario could be made reality by a neuroscientist who continually measures whether B is actualised. In this case, you will come to the erroneous conclusion that raising your hand causes$_1$ the beep. But such scenarios are not common. Our brain states

[3] From an anthropological perspective, one can speculate that animism—the ascription of personhood to natural objects—is the temporary result of such a development, where causation has been placed outside of human actions, but has not yet lost its connection to action completely. We are not competent to develop this suggestion.

tend to have few noticeable effects except through our actions, and the correlations between prior events outside of our brain and our actions, while they do exist, tend to disappear when we decide to break our regular patterns of behaviour.

For example, the rising of the sun may cause me to get out of my bed, and it may cause the singing of the birds. This is a common cause structure that could lead me astray and make me conclude that my rising causes the birds to sing. But as soon as I try, whenever the whim takes me, to make the birds sing by getting out of bed, the correlation disappears. And this happens to be the case for almost every common cause structure involving actions that is instantiated in our universe. Our justified beliefs about causal$_1$ relations will therefore almost always be held to be true causal$_2$ claims as well, and this ensures that they can be used as the basis of a bootstrapping procedure.

We now return to the question of why a generalisation of causation$_1$ to non-action causes would go towards an interventionist notion. A naïve generalisation, which drops the action requirement and just looks at correlations, will have a problem with common cause structures: as is well known, correlations between X and Y do not prove that X causes Y. Since common cause structures, which happen to be almost always absent between actions and their seeming results, are common in the rest of nature, this is a major problem.

To solve this problem, we need to require that there exists a variable I which can change some non-action variable X independently of other variables in the environment. We then observe whether such 'action-like' changes of X correlate with changes in Y, and if they do, we conclude that X causes$_2$ Y. To make sure that we do not fall prey to common cause structures and other potential problems, we need to exclude several possibilities: e.g., I influencing Y through a causal$_2$ path that does not go through X, or I being correlated with a variable Z that influences Y through a causal$_2$ path that does not go through X. Such possibilities are exactly what Woodward attempts to exclude with his idea of intervention variables (Woodward 2003a, p. 98).

We know from our previous discussions that to know that a given variable is a Woodwardian intervention variable, we need to use prior causal$_2$ knowledge. If we start with an agency theory, we have prior causal$_1$ knowledge. Transposing that knowledge to the new theory of causation$_2$, we have a basis from which we can develop further causal$_2$ knowledge of the world. This is possible because our causal$_1$ knowledge is almost entirely free of common cause structures. It is plausible because the notion of causation$_2$ can be seen as a more sophisticated development of the notion of causation$_1$. We have bootstrapped ourselves out of the swamp of agency into the impersonal realm of interventionism.

In this section, then, we have seen why agents might develop the notion of causation$_1$; how they can get justified beliefs about causal$_1$ relations; why causation$_1$ must be generalised to non-actions; why the problems with such a generalisation will lead towards a Woodwardian notion of causation$_2$; and that our world is in fact such that if we turn our causal$_1$ beliefs into causal$_2$ beliefs, most of them can be held true, which proves that they can function as a bootstrapping basis. This amounts to a rational reconstruction of the genesis of our causal$_2$ knowledge, and therefore solves the problem of genesis that besets Woodward's theory. To seal the deal, so to speak, we will now argue that this story of genesis lines up well with our empirical knowledge of the development of the concept of causation.

4 Agency and interventionism from an empirical perspective

In the previous sections we argued that the problem of circularity for Woodward's interventionism can be solved if we assume that the interventionist notion of causation can be understood as a sophisticated version of an agency notion. More in particular, we claimed that the relation between interventionism and agency is strong enough to enable a bootstrap procedure from the latter to the former. In this section we show how such a bootstrap procedure can be explicated in empirical terms by distinguishing between four stages of development:

(i) perceiving correlations between non-action events (i.e., perceiving temporal and spatial contiguity);
(ii) understanding causal$_1$ relations between one's own actions and their effects;
(iii) understanding causal$_1$ relations between another agent's actions and their effects;
(iv) understanding causal$_2$ relations between non-action events.

Before spelling out these different stages in more detail, we would like to stress that we do not attempt to provide a full-fledged ontogenetic model of the interventionist notion of causation we have presented so far. Our aim is much more modest; it is to establish that, if one considers the development of causal understanding from a naturalistic perspective, there are interesting facts that should encourage us to take this notion seriously. Furthermore, we do not want to suggest that this development can be neatly divided in four stages, each with a precise onset time. Although we will provide an indication of the onset of each stage, based on what is currently known about the development of causal understanding, this is of course constrained by various experimental limitations and subject to further empirical testing.

Empirical research on causal understanding in early infancy has primarily focused on the capacity to perceive temporal and spatial contiguity between events—what Leslie (1995) referred to as 'mechanical causality'. This Humean notion of causality has been tested by means of 'direct launching' experiments (Michotte 1963), in which subjects are presented with a stimulus X that moves across the screen and makes contact with a second stimulus Y that then begins to move. Subjects perceive this event as 'causal', but they fail to do so if (1) a temporal delay is inserted from the time the stimuli make contact until the second stimulus begins to move, or (2) a spatial distance is created between the two stimuli. Leslie and Keeble (1987) used this paradigm to show that the ability to perceive causal relations in terms of temporal and spatial contiguity is already present in 6-month-olds (see also Schlottmann and Surian 1999 for similar findings in 9-month-olds).

The capacity to perceive temporal and spatial contiguity between events provides agents with a basic understanding of predictive relations, e.g. that stimulus X predicts stimulus Y. This understanding is still limited and not genuinely causal; insofar as the predictive relations hold for events that are outside of the agent's control, there is no way in which the agent can know whether stimulus X actually causes stimulus Y.

Things are different, however, when we consider the relations between the agent's *own actions* and their effects. As we argued in Sect. 3, the agent can easily test whether her action X is the cause of effect Y by doing X in circumstances in which she otherwise would not do X, and see if this does or does not break the correlation with Y

(or refrain from doing X, and see whether Y still happens). The only requirement for such a testing procedure is voluntary action control. We can use Elsner and Hommel's (2001) two-stage model of voluntary action control to explain how agents acquire an understanding of causal$_1$ relations between their own actions and their effects (see also Hommel and Elsner 2009). Suppose an agent performs a particular action *X* that leads to a particular effect *Y*. According to Elsner and Hommel's model, the motor and perceptual representations underlying respectively the action X and the perception of effect Y are integrated in such a way that activating the perceptual representation of effect Y on a later occasion also activates the motor representation of action X. In this way, the agent learns that action X has to be performed in order to bring about effect Y.[4] Empirical studies indicate that this primitive agency notion of causation emerges around 9 months of age (e.g., Verschoor et al. 2010).

The agent's understanding of causal$_1$ relations between her own actions and their effects is intimately intertwined with her understanding of causal$_1$ relations between *another* agent's actions and their effects. For example, Sommerville et al. (2008) have shown that 10-month-old infants who received active training in pulling a cane to retrieve a toy learned about the causal relation between another agent's cane-pulling action and toy retrieval more readily than those who relied on observational training. These findings have been taken to show that action and perception are essentially coupled and share the same 'representational space' (Georgieff and Jeannerod 1998). Further support for this assumption comes from, the discovery of so-called 'mirror mechanisms' in the brain (Rizzolati and Craighero 2004; Gallese and Sinigaglia 2011), developmental studies on imitation (Meltzoff 2004, 2006; Meltzoff and Moore 1977, 1994; Meltzoff and Brooks 2001), and considerations about representational formats (Kovács et al. 2010).

How do agents acquire an understanding of causal$_2$ relations between non-action events? Woodward (2007) suggests that this requires a full-fledged means/ends understanding, and an appreciation of causal relations among variables that are intermediate between the agent's action and its effect.

> In particular, means/ends understanding seems to involve a decomposition of a task into an intermediate outcome O that can be produced fairly directly by the subject's action A and a further outcome O' that is more directly caused by O and less directly by A, and where the link between O and O' is a tertiary link between events, rather than an action-event link. (p. 34)

According to Woodward, the postulation of such an intermediate link goes hand in hand with a decoupling of means from ends and a focus on the latter as a separate entity. Following Tomasello and Call (1997), he argues that this decoupling is closely

[4] The difference between perceiving correlations between external events (stage i) and understanding causal$_1$ relations between another agent's actions and their effects (stage ii) resembles the difference between classical and instrumental conditioning. In classical conditioning, the agent learns about a predictive relation between two events that are outside of her control, whereas in instrumental conditioning what is learned is a predictive relation between an action of the agent and its effect. Woodward (2007) argues that, from an interventionist perspective, instrumental learning has a 'cause-like' flavor. Although we agree that instrumental learning has this cause-like flavor, we think this is better explained using an agency perspective on causation.

linked to learning through imitation, where the behavioral means have to be copied in such a way as to reproduce the goal of the action, while other features have to be varied in order to accommodate differences between the target and the imitator.

Although 6-month-olds are already capable of reproducing another agent's action on an object (Barr et al. 1996), imitative learning in the sense described above seems to emerge only after the first year of life. In an experiment by Meltzoff (1988), for example, 14-month-olds observed an experimenter bend down and activate a rectangular box with his head, causing the box to light up. The infants followed suit even though they were in the position to turn on the light by simpler means (e.g., with their hands), which means that they were reproducing the means-end structure of the action, and not just the goal (see also Gergely et al. 2002; Williamson and Markman 2006).

In this experiment, the link between the intermediate outcome (activating the rectangular box) and the further outcome (turning on the light) is still an action-event link. However, their increasing flexibility in reasoning about means (i.e. that goals can be brought about in very different ways) eventually allows infants to understand an intermediate outcome as a non-action event, and hence the link between an intermediate outcome and a further outcome as a $causal_2$ relation. We are not aware of any empirical evidence that shows precisely when this happens. As Bonawitz et al. (2010) have shown, young infants (24 months) still need supplemental information to understand $causal_2$ relations between non-action events. That is, infants do not spontaneously intervene on a predictive relation unless the events are initiated by another agent, the events involve unmediated, direct contact between objects, or adults describe the events in causal language. However, Bonawitz et al. (2010) found that older infants (from 37 months onwards) do spontaneously intervene without this kind of information.

Many loose ends remain in the outline sketched above. What we hope to have shown, however, is that there is a plausible empirical story to tell about the acquisition of an interventionist understanding of causality in the early life of the child, one that lines up with the two-stage rational reconstruction we presented in Sect. 3. This gives psychological plausibility to our philosophical theory. In addition, our story seems to be compatible with Woodward's own empirical account of interventionism (as presented in Woodward 2007, 2009), especially since he acknowledges that agency-based accounts may play an important role in the acquisition of causal knowledge (see Woodward 2009, p. 259).

5 Conclusion

Woodward's theory is beset by a problem of circularity. This circularity should be seen, not as a problem in Woodward's analysis of the meaning of causation, but as a problem of genesis: on Woodward's theory, it is mysterious that we ever manage to start getting causal knowledge. Interventionism may work very well once we have a basis of causal knowledge, but it cannot explain how we ever get such a basis.

We have presented a way out of this dilemma by showing how a primitive agency notion of causation can be used to get the interventionist theory of causation started. The concept of intervention can be understood as a generalisation of the concept of agency by which the limitations and problems of the more primitive theory can be

overcome. At the same time, it (contingently) turns out that the positive causal claims of the agency theory can almost always be held true on the interventionist theory, which means that they can be used as a basis for the more advanced theory.

We have also shown that our rational reconstruction of the development from agency to intervention lines up nicely with empirical research about causal reasoning in children. This lends further plausibility to our philosophical theory.

Finally, it should be stressed again that the theory presented here is not a hybrid agency/interventionist theory. The analysis of causation given by Woodward remains wholly intact. We have merely sketched a path for the development of this concept of causation, a development which, because of the circularity of the concept, was mysterious. By removing the mystery, we believe to have strengthened the interventionist position considerably.

Acknowledgments We would like to thank two anonymous referees, Lena Kästner, and Markus Eronen for their valuable comments and suggestions. We would also like to thank the audience of the Zentrum für interdisziplinäre Forschung workshop 'Agents and Causes', and the HPS group of Leiden University's Institute of Philosophy, for their feedback on a related paper presentation.

References

Barr, R., Dowden, A., & Hayne, H. (1996). Developmental changes in deferred imitation by 6- to 24-month-old infants. *Infant Behavior and Development, 19*, 159–170.

Baumgartner, M. (2009). Interdefining causation and intervention. *Dialectica*, *63*, 175–194.

Beebee, H., Hitchcock, C., & Menzies, P. (2009). *The Oxford handbook of causation*. New York: Oxford University Press.

Bonawitz, E., Ferranti, E., Saxe, R., Gopnik, A., Meltzoff, A. N., Woodward, J., et al. (2010). Just do it? Investigating the gap between prediction and action in toddlers' causal inferences. *Cognition*, *115*, 104–117.

Cartwright, N. (1994). *Nature's capacities and their measurement*. New York: Oxford University Press.

Collingwood, P. G. (1940). *An essay on metaphysics*. Oxford: Clarendon Press.

de Regt, H. (2004). Review of James Woodward, making things happen. *Notre Dame Philosophical Reviews*. http://ndpr.nd.edu/news/23818-making-things-happen-a-theory-of-causal-explanation/

Elsner, B., & Hommel, B. (2001). Effect anticipation and action control. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 229–240.

Gallese, V., & Sinigaglia, C. (2011). What is so special about embodied simulation? *Trends in Cognitive Sciences*, *15*(11), 512–519.

Georgieff, N., & Jeannerod, M. (1998). Beyond consciousness of external reality: A "Who" system for consciousness and action and self-consciousness. *Consciousness & Cognition*, *7*, 465–487.

Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*, 755.

Glymour, C. (2004). Critical notice of James Woodward, making things happen. *British Journal for the Philosophy of Science*, *55*, 779–790.

Hommel, B., & Elsner, B. (2009). Acquisition, representation, and control of action. In E. Morsella, J. A. Bargh, & P. M. Gollwitzer (Eds.), *Oxford handbook of human action* (pp. 371–398). New York: Oxford University Press.

Kovács, A., Teglas, E., & Endress, A. (2010). The social sense: Susceptibility to others' beliefs in human infants and adults. *Science*, *330*, 1830–1834.

Leslie, A. M. (1995). A theory of agency. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition* (pp. 121–141). Oxford: Clarendon Press.

Leslie, A. M., & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition*, *25*, 265–288.

Meltzoff, A. N. (1988). Infant imitation after a 1-week delay: Long-term memory for novel acts and multiple stimuli. *Developmental Psychology*, *24*, 470–476.

Meltzoff, A. N. (2004). Imitation and other minds: The "like me" hypothesis. In S. Hurley & N. Chater (Eds.), *Perspectives on imitation: From neuroscience to social science* (Vol. II, pp. 55–77). Cambridge, MA: MIT Press.

Meltzoff, A. N. (2006). The "like me" framework for recognizing and becoming an intentional agent. *Acta Psychologica*, *124*, 26–43.

Meltzoff, A. N., & Brooks, R. (2001). 'Like me' as a building block for understanding other minds: Bodily acts, attention, and intention. In B. F. Malle, L. J. Moses, & D. A. Baldwin (Eds.), *Intentions and intentionality: Foundations of social cognition* (pp. 171–191). Cambridge, MA: MIT Press.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science*, *198*, 75–78.

Meltzoff, A. N., & Moore, M. K. (1994). Imitation, memory, and the representation of persons. *Infant Behavior and Development*, *17*, 83–99.

Menzies, P., & Price, H. (1993). Causation as a secondary quality. *British Journal for the Philosophy of Science*, *44*, 187–203.

Michotte, A. E. (1963). *The perception of causality* (T. R. Miles & E. Miles, Trans.). London: Methuen (original work published 1946).

Psillos, S. (2002). *Causation & explanation*. Montreal: McGill-Queen's University Press.

Rizzolati, G., & Craighero, L. (2004). The mirror–neuron system. *Annual Review of Neuroscience*, *27*, 169–192.

Schlottmann, A., & Surian, L. (1999). Do 9-month-olds perceive causation-at-a-distance? *Perception*, *28*, 1105–1113.

Sommerville, J. A., Hildebrand, E. A., & Crane, C. C. (2008). Experience matters: The impact of doing versus watching on infants' subsequent perception of tool use events. *Developmental Psychology*, *44*, 1249–1256.

Sosa, E., & Tooley, M. (1993). *Causation*. New York: Oxford University Press.

Tomasello, M., & Call, J. (1997). *Primate cognition*. New York: Oxford University Press.

Verschoor, S. A., Weidema, M., Biro, S., & Hommel, B. (2010). Where do action goals come from? Evidence for spontaneous action-effect binding in infants. *Frontiers in Psychology*, *1*, 201.

Von Wright, G. (1971). *Explanation and understanding*. Ithaca, NY: Cornell University Press.

Williamson, R. A., & Markman, E. M. (2006). Precision of imitation as a function of preschoolers' understanding of the goal of the demonstration. *Developmental Psychology*, *42*, 723–731.

Woodward, A. L. (2003a). Infants' developing understanding of the link between looker and object. *Developmental Science*, *6*, 297–311.

Woodward, J. (2003b). *Making things happen*. New York: Oxford University Press.

Woodward, J. (2007). Interventionist theories of causation in psychological perspective. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy and computation* (pp. 19–36). New York: Oxford University Press.

Woodward, J. (2008a). Causation and manipulability. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy (winter 2008 edition)*. http://plato.stanford.edu/archives/win2008/entries/causation-mani/.

Woodward, J. (2008b). Response to Strevens. *Philosophy and Phenomenological Research*, *LXXVII*, 193–212.

Woodward, J. (2009). Agency and interventionist theories. In H. Beebee, C. Hitchcock, & P. Menzies, P (Eds.), *The Oxford handbook of causation* (pp. 234–262). New York: Oxford University Press.