



Universiteit  
Leiden  
The Netherlands

## Explanation and determination

Gijsbers, V.A.

### Citation

Gijsbers, V. A. (2011, August 28). *Explanation and determination*. Retrieved from <https://hdl.handle.net/1887/17879>

Version: Corrected Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/17879>

**Note:** To cite this publication please use the final published version (if applicable).

## Chapter 4

# A General Interventionist Theory of Explanation: Extending Woodward's Account

### 4.1 Introduction

Recent efforts in the construction of theories of explanation have been divided between general theories, which aim to capture the essence of all explanations, and domain-restricted theories, which focus on specific kinds of explanations. The dominant general theory is the unificationist theory, which has been defended in Kitcher 1981 [55] and 1989 [57], Schurz & Lambert 1994 [120], and Schurz 1999 [119]. The most important domain-restricted theory is the theory of causal explanation defended in Salmon 1984 [110] and 1998 [113], Woodward 2003 [142] and Strevens 2008 [132].

General theories are preferable to domain-restricted theories, since they give us more information about what explanation actually is. On the other hand, they are also more difficult to get right, since they have to be true about a greater range of instances. If our general theories are not up to this task, as Gijsbers 2007 [31] argues about the unificationist theory, we may have to settle temporarily for the less ambitious domain-restricted theories. But the philosopher's task would nevertheless remain to search for a truly general theory of explanation; and I hope to contribute to that search with the present chapter.

One way to approach this task is by taking a successful domain-restricted theory and generalising it so that it encompasses kinds of explanation that

lie outside its original domain. In this chapter, I will attempt such a project. Taking James Woodward's interventionist theory of causal explanation, which he has laid down in his book *Making Things Happen* [142], as a starting point, I will attempt to show that its basic ideas can be generalised to encompass non-causal explanations.

In section 4.2, I will give some examples of non-causal explanations, and briefly comment on the way in which one could generalise a theory of causal explanation to a theory of explanation *simpliciter*. Woodward's theory is discussed in section 4.3, and defended against the charge of circularity in section 4.4.

After the domain-restricted theory has been elucidated, we will start extending it by applying it to a specific set of non-causal explanations: most extensively to mathematical explanations in sections 4.5, 4.6 and 4.7. The aim here is to show that in the example discussed, a Woodwardian analysis yields the right judgments about mathematical explanation. We then tackle the explanation of laws by more general laws in section 4.8. The lessons learned in these sections are used in section 4.9 to formulate a generalised interventionist theory of explanation.

## 4.2 Causality and explanation

Not all explanations are causal. To give some examples of non-causal explanations:

1. Mathematical explanations that show how one mathematical fact follows from and is explained by others. According to most philosophies of mathematics, mathematical facts cannot enter into causal relations.
2. Explanations of laws of nature by other laws of nature – for instance, an explanation of Kepler's Laws by showing that they follow from Newton's Laws and certain initial conditions. Since laws are not events in time and space, they cannot be said to *cause* each other.
3. Explanations of human actions by desires, beliefs and decisions of the agent – elements which cannot enter into causal relations according to some (though by no means all) philosophers.
4. Transcendental explanations of the kind proposed in Kant 1787 [54]. According to Kant, that we see chains of cause and effect in the world is to be explained by the fact that cause is one of the categories inherent in our mind. It would be meaningless to say that this category *causes* the world to be causal.

In what follows, I will discuss only the first two of these examples, since they are the least controversial. Whether or not actions are part of the network of cause and effect is a contentious topic in the philosophy of action; and it has also been claimed that transcendental explanations only *seem* to furnish understanding, and seem to do this because we cannot help but interpret them as causal mechanisms.<sup>1</sup> Discussing these topics would take us too far afield.

However, although not all explanations are causal, there is nevertheless a strong link between explanation and causality. This is immediately clear when we read works that present theories of causal explanation. The majority of these works, and Woodward's is no exception (though Strevens's is), spend much more time and effort defining *causality* than defining *causal explanation*: the link between the two is taken to be so strong that once you know what a cause is, explicating the concept of explanation becomes trivial, something along the lines of "to explain *X* you must give a cause of *X*". In generalising a theory of causal explanation to non-causal explanations, we must therefore generalise the notion of *cause*, for without that notion, there is little to generalise.

One more remark before we go on to Woodward's analysis of the notions of cause and causal explanation. When it comes to giving a theory of causality, philosophers can embark on two very different projects, which Dowe (Dowe 2000, [25] pp. 1-13) calls conceptual and empirical theories of causality. (An analogous distinction is discussed in a more general way in Millikan 1989 [82] and Neander 1991 [88] under the names 'conceptual analysis' and 'theoretical definition'.) By linking causation with explanation, we are taking the path of a conceptual theory; and we will leave empirical theories like Dowe's conserved quantity theory to one side. That these two projects are different can be easily seen: while it is obvious that the notion of 'conserved quantity' will not feature in a satisfactory analysis of our concept of explanation, it may nevertheless be at the core of a satisfactory *empirical* theory of causality (see Dowe 1992 [23], 1995 [24], 2000 [25]); and *idem dito* for any other empirically discovered feature of the actual causal processes in the world.

### 4.3 Woodward on causation and explanation

I will now present James Woodward's interventionist theory of causal explanation, starting with his analysis of causation.

---

<sup>1</sup>Rorty 1979 [106], p. 151; especially footnote 31.

Woodward takes as primitives variables ranging over (type-level) events. We might, for instance, have a variable  $W$  which ranges over  $\{\text{'it rains'}, \text{'it does not rain'}\}$ . In a system  $\mathbf{V}$  of such variables Woodward defines *direct causes* and *contributing causes* in terms of probabilities and interventions:

A necessary and sufficient condition for  $X$  to be a (type-level) direct cause of  $Y$  with respect to a variable set  $\mathbf{V}$  is that there be a possible intervention on  $X$  [with respect to  $Y$ ] that will change  $Y$  or the probability distribution of  $Y$  when one holds fixed at some value all other variables  $Z_i$  in  $\mathbf{V}$ . A necessary and sufficient condition for  $X$  to be a (type-level) *contributing cause* of  $Y$  with respect to a variable set  $\mathbf{V}$  is that (i) there be a directed path from  $X$  to  $Y$  such that each link in this path is a directed causal relationship; that is, a set of variables  $Z_1 \dots Z_n$  such that  $X$  is a direct cause of  $Z_1$ , which is in turn a direct cause of  $Z_2$ , which is a direct cause of  $\dots Z_n$ , which is a direct cause of  $Y$ , and that (ii) there be some intervention on  $X$  [with respect to  $Y$ ] that will change  $Y$  when all other variables in  $\mathbf{V}$  that are not on this path are fixed at some value.<sup>2</sup> ([142], p. 59.)

Because causes are defined relative to a variable set  $\mathbf{V}$ , causation in Woodward's theory is description-dependent. Let us give an example. Sending funds to third-world countries ( $M$ ) tends to increase development ( $D$ ), but also to increase corruption ( $C$ ), which in turn has a negative effect on development. Suppose that the good and the bad cancel each other exactly if there are no other interventions. Now, if  $\mathbf{V}$  contains  $M$ ,  $C$  and  $D$ , Woodward's theory will have us conclude that  $M$  is a direct cause of  $D$ , because if we hold  $C$  fixed (perhaps by political reforms or pressure), changing  $M$  will change  $D$ . But if, on the other hand,  $\mathbf{V}$  contains only  $M$  and  $D$ , the theory will have us conclude that  $M$  is *not* a direct (or even indirect) cause of  $D$ .

We turn now to the definition of 'intervention', which is obviously a central term in the interventionist theory of causal explanation. Woodward characterises intervention in two steps. First, the notion of 'intervention variable' is defined.  $I$  is an intervention variable for  $X$  with respect to  $Y$  if and only if  $I$  meets the following conditions:

- I1.  $I$  causes  $X$ .

---

<sup>2</sup>A redundant part of the definition has been left out. The phrases between square brackets have been added to clear up the relation of these definitions with the definition of intervention below.

- 12.  $I$  acts as a switch for all the other variables that cause  $X$ . That is, certain values of  $I$  are such that when  $I$  attains those values,  $X$  ceases to depend on the values of other variables that cause  $X$  and instead depends only on the value taken by  $I$ .
- 13. Any directed path from  $I$  to  $Y$  goes through  $X$ . [...]
- 14.  $I$  is (statistically) independent of any variable  $Z$  that causes  $Y$  and that is on a directed path that does not go through  $X$ .<sup>3</sup> ([142], p. 98.)

(We will talk about the cause-intervention circularity established by these definitions in the next section.) Then, intervention is defined as follows:

$I$ 's assuming some value  $I = z_i$ , is an intervention on  $X$  with respect to  $Y$  if and only if  $I$  is an intervention variable for  $X$  with respect to  $Y$  and  $I = z_i$  is an actual cause of the value taken by  $X$ . ([142], p. 98.)

Before discussing this any further, I will give Woodward's definition of explanation.

Suppose that  $M$  is an explanandum consisting in the statement that some variable  $Y$  takes the particular value  $y$ . Then an explanans  $E$  for  $M$  will consist of (a) a generalization  $G$  relating changes in the value(s) of a variable  $X$  (where  $X$  may itself be a vector or n-tuple of variables  $X_i$ ) and changes in  $Y$ , and (b) a statement (of initial or boundary conditions) that the variable  $X$  takes the particular value  $x$ . A necessary and sufficient condition for  $E$  to be (minimally) explanatory with respect to  $M$  is that (i)  $E$  and  $M$  be true or approximately so; (ii) according to  $G$ ,  $Y$  takes the value  $y$  under an intervention in which  $X$  takes the value  $x$ ; (iii) there is some intervention that changes the value of  $X$  from  $x$  to  $x'$  where  $x \neq x'$ , with  $G$  correctly describing the value  $y'$  that  $Y$  would assume under this intervention, where  $y' \neq y$ . ([142], p. 203.)

Given the relation between causation and intervention, this is (give or take a few niceties) equivalent to the following claim: an explanation of  $Y = y$  consists of a statement  $X = x$  and a true story which shows that  $X$

---

<sup>3</sup>A redundant part of the definition has been left out.

is a (type-level) cause of  $Y$ , and how  $Y$  depends on  $X$ . As I said in section 4.2, once we have defined causation, it is easy to get to explanation.

One of the niceties that is worth commenting upon is the word ‘minimally’ in Woodward’s definition of explanation. According to Woodward, explanation is not an all-or-nothing affair. We can have minimal explanations (which give us relatively little understanding), and fuller explanations (which give us more understanding). A good explanation doesn’t just conform to the above definition, but will “involve a generalization  $G$  and explanans variable(s)  $X$  such that  $G$  correctly describes how the value of  $Y$  would change under interventions that produce a range of different values of  $X$  in different background circumstances” (Woodward 2003 [142], p. 203). As an example, let  $G_1$  be “these balls  $A$  and  $B$  will both lie still after a collision if and only if they have opposite velocities when they collide”, and let  $G_2$  be “two inelastic balls will both lie still after a collision if and only if they have opposite momenta, which is velocity times mass, when they collide”. The second of these generalisations allows for better explanations than the first, for it allows us to predict what will happen in more circumstances. Common ground is touched here with the unificationist theory, but note that while in Woodward’s theory unificatory power helps to make an explanation a better explanation, it is not a *conditio sine qua non* for being an explanation.

## 4.4 Is the interventionist theory circular?

I will not defend Woodward’s theory of causality and explanation here – the reader is urged to consult his book – except in one respect, which is the problem of circularity. Woodward’s theory is *prima facie* viciously circular: causality is defined using the notion of intervention, and intervention is defined using the notion of causality. Of course Woodward anticipates this criticism and offers two counterarguments (Woodward 2003 [142], pp. 104–107). Some reviewers agree with these counterarguments (Weslake 2006 [138]; Menzies 2006 [81]), although sometimes with reservations (Markus 2007 [77]; Hiddleston 2005 [45]); but at least two reviewers are not convinced and argue that the circularity of Woodward’s theory makes it fail as an analysis of causation (de Regt 2004b [101]; Glymour 2004 [33]). This problem is too important for me to ignore. I will first give Woodward’s two arguments; then I will summarise the counterarguments of De Regt and Glymour; and finally, I will sketch the outline of a counter-counterargument.

In his first argument, Woodward casts doubt upon the possibility of a reductive theory of causality, that is, a theory of causality that reduces causal claims to claims formulated entirely in a non-causal vocabulary. Speaking of

Humean and correlation-based theories, Woodward asserts that “the failure of this sort of reduction has been a familiar theme in recent philosophical discussion”. For instance, information about the correlations between  $X$ ,  $Y$  and  $Z$  underdetermines whether  $X$  is a common cause of  $Y$  and  $Z$  or whether  $Y$  is a cause of  $X$  which in turn is a cause of  $Z$ . Causality, according to Woodward, goes beyond the categories of a sparse empiricism, and a reductive analysis is therefore not to be expected. If no reductive analysis is possible, we will have to make do with a non-reductive one.

In his second argument, Woodward points out that on his definitions of causation and intervention, we must have *some* causal knowledge in order to decide whether  $X$  causes  $Y$ , but what we do not need to know in advance is whether  $X$  causes  $Y$ . Therefore, although the theory is non-reductive, it is not viciously circular: you don’t have to know something before you know it.

These arguments do not convince Henk de Regt and Clark Glymour. De Regt explicitly argues against them, writing:

MT [Woodward’s theory about the meaning of causal claim] 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. (De Regt 2004 [101].)

Glymour grants Woodward that his theory is not circular, but he is nevertheless in substantial agreement with De Regt:

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’. (Glymour 2004 [33], p. 785.)

In both cases, the argument is that Woodward’s theory can be used only if we already know what causality is, or already have a bunch of things we know are direct causes: Woodward’s theory cannot explain how causal discourse

ever got going. Therefore, it cannot be a satisfactory theory of the meaning of causality.

Both De Regt and Glymour presuppose that a satisfactory analysis of the meaning of a term must tell us how we can introduce the predicate into a language which does not already contain it; we must specify the meaning of the new predicate entirely in terms of the old predicates. On such a view, a predicate is either primitive and hence unanalysable, or non-primitive and completely analysable into other (primitive or at least ‘more’ primitive) concepts. If a satisfactory analysis of ‘cause’ exists at all, it cannot contain the predicate ‘cause’; so something is wrong with Woodward’s analysis.

But is it true that the only ways to introduce a new predicate into a language is either by introducing it as a primitive (however that may work), or by giving a reductive analysis in terms of other predicates? Let us draw an analogy between Woodward’s analysis of causality and Wilfrid Sellars’s classic defence of non-reductive analyses of mentalistic concepts in *Empiricism and the Philosophy of Mind* (Sellars 1997 [122]). Suppose, Sellars says, that it is clear how we acquire concepts that refer to publicly accessible states of affairs; then there is still the problem of how we acquire concepts that refer to ‘private’ states of affairs, such as mental states; for how would the criteria of correct use be established? Of course, this problem would immediately be solved if we could find a reductive analysis of mentalistic concepts, but according to Sellars (pp. 86-117) this is not possible. So how *did* we get mentalistic concepts in our language?

Sellars uses a thought experiment (the ‘myth of Jones’) to show that these concepts may first have been introduced into our language as theoretical entities that helped explain and predict people’s overt behaviour. (In exactly the same sense concepts like ‘electron’ are introduced into our language as theoretical entities that help explain and predict the overt behaviour of physical systems.) But theoretical discourse can go on to lead a life of its own. In the case of mental concepts, Sellars claims that “it now turns out – need it have? – that Dick can be trained to give reasonably reliable self-descriptions, using the language of the theory, without having to observe his overt behaviour. [...] *What began as a language with a purely theoretical use has gained a reporting role.*” (Sellars 1997 [122], pp. 106-107). What happens is that terms like ‘thought’ and ‘angry’ are introduced in a reductive fashion. It then turns out to be the case, empirically, that people can reliably tell that they are angry even before this anger has manifested itself; which leads us to adopt a new and wider meaning of ‘anger’ that can even apply in circumstances when the anger does *not* manifest itself at all. “He is angry but he does not show it” used to be a contradiction, but now becomes accepted usage. Once this happens, talk of inner episodes is no longer reducible to that of overt

behaviour.

What Sellars argues in the case of mentalistic concepts is that although they could be introduced as theoretical terms useful for explaining overt behaviour, it turned out to be the case that they could also be used to report on inner episodes. This new role was irreducible to the kind of language which was originally used to give them meaning; hence, mentalistic concepts cannot be reductively analysed.

I would like to suggest that a similar history can be written for the notion of ‘cause’, although here the problem is not to bridge the gap between the public and the private, but between correlation and determination. It is clear that causal notions could have been introduced as theoretical terms that help us to predict and explain ‘overt’ concepts like constant conjunction and correlation. It then *turned out* to be the case that they could also be used to explain, plan and predict the success of our own interventions in nature. In these instances, the experience of agency showed cause and effect to be something other than mere correlation. After the concept of cause had been given this new interpretation in instances of agency, it could be projected back on the natural world. In its new role, it had become irreducible to a sparse empiricist vocabulary.<sup>4</sup>

An analysis of the concept of cause could then consist of two steps: we first introduce some instances of correct usage of ‘cause’ by appealing to the experience of agency; and then we give the empirico-semantic rules that allow us to generate more (and hopefully all) instances of correct usage from the basis given in the first step. I take Woodward’s analysis of causality to be precisely this second step: it shows us how to extend the notion of cause beyond the immediately graspable experience of agency. Woodward does not talk about the first step, presumably because he wishes to distance himself from charges of subjectivism, but this leaves him open to the charges of Glymour and De Regt. Without the first step, it remains mysterious how we ever get causal talk going.

We do not need to be worried about appealing to subjective experience in our analysis of ‘cause’. Any subjectivism that remains after the completion of the second step, where the notion of ‘cause’ is embedded into an inter-subjective language of public states of affairs, is harmless. So Woodward is right that it is intervention rather than agency that should lie at the core of analyses of causality; but Weslake (2006 [138]) is right that Woodward should take the notion of agency more seriously than he currently does. We need agency to get causal discourse starting; but once it has started, Wood-

---

<sup>4</sup>This understanding of causation through agency is closely related to Schopenhauer’s use of the will to explain the notion of ground; see Schopenhauer 1813 [117].

ward shows us how to objectivise the notion of cause so that we end up not needing agency any more.

Much remains to be said about this topic, since there are many deep and interesting questions both about what makes a good analysis and about the experience of agency. But I hope that my sketchy treatment of the issue has at least shown that there is ground on which Woodward can fight against charges of circularity. I will now return to the main line of my argument.

## 4.5 Mathematical explanation introduced

Mathematical explanations are non-causal according to all but the most unconventional philosophies of mathematics, and they will therefore serve as the prime example of this chapter.<sup>5</sup> I wish to take an example of a mathematical explanation and a mathematical non-explanation, and show that a generalised version of Woodward's theory correctly implies that the former explains and the latter does not.

Now compared to the amount that has been written about explanation in the sciences, very little has been written by philosophers on the subject of mathematical explanation; important exceptions are Steiner 1978 [127], Sandborg 1998 [116], Mancosu 2001 [76] and Hafner and Mancosu 2005 [36], as well as parts of Kitcher 1989 [57]. This lack of attention means that there are no standard examples of mathematical explanation (like the barometer, the flagpole, the birth control pills or the double assassination in other areas of the philosophy of explanation) with which my readers will be certainly familiar. The examples to be used must therefore first be set out in some detail; and I will do so in the rest of this section, using an example that has received some discussion in the extant literature. In section 4.6, I will go on to discuss Steiner's theory of mathematical explanation, and his attempt to make sense of the examples. Steiner's theory serves as a natural prelude to testing a generalised version of Woodward's theory in section 4.7.

Mathematicians are in the business of proving theorems. But different proofs of a theorem, although they all (by virtue of being proofs) give us certainty *that* the theorem is true, may not give us equal understanding of *why* the theorem is true. That is, some proofs explain the theorem they prove, and others do not.<sup>6</sup> Mathematicians find this distinction important, and will

---

<sup>5</sup>We will look only at mathematical explanations that take the form of proofs. There are probably other kinds of mathematical explanation – such as explanations by pictures – but even less work has been done on them.

<sup>6</sup>I will argue in section 4.7 that the proof cannot strictly speaking be identified with the explanation.

often continue to search for an explanatory proof even if a non-explanatory proof is already available. (See Mancosu 2001 [76] and especially Hafner and Mancosu 2005 [36] for evidence that this is indeed the case.)

Let us consider an elementary theorem and two proofs, one explanatory and one not explanatory. This example is taken from Steiner 1978 [127].

**Theorem:** Let  $S(n)$  be  $1 + 2 + \dots + n$ . Then  $S(n) = \frac{n(n+1)}{2}$ .

**First proof:** This is a proof by induction. We start by showing that *if* the theorem holds for  $n$ , *then* it holds for  $n + 1$ :

$$S(n+1) = S(n) + (n+1) = \frac{n(n+1)}{2} + \frac{2(n+1)}{2} = \frac{(n+1)(n+2)}{2}. \quad (4.1)$$

The theorem holds for  $n = 1$ :  $\frac{1(1+1)}{2} = 1$ . Therefore, by complete induction over the natural numbers, the theorem follows.

**Second proof:** By the commutativity of addition,  $S(n) = 1 + 2 + \dots + n$  is equal to  $S(n) = n + (n-1) + \dots + 1$ , the same sum but reversed. We now add the rewritten to the original sequence term by term:

$$2S(n) = [1 + 2 + \dots + n] + [n + (n-1) + \dots + 1] \quad (4.2)$$

$$= [1 + n] + [2 + (n-1)] + \dots + [n + 1] \quad (4.3)$$

$$= [n + 1] + [n + 1] + \dots + [n + 1] \quad (4.4)$$

$$= n(n+1). \quad (4.5)$$

The theorem follows immediately.

Steiner claims, and I agree, that unlike the first proof, the second proof not merely proves, but also explains the result. We might almost be tempted to say that in the second proof, we can see how the symmetry of the sequence 'causes' the sum to be what it is. Of course, there is no causation; but what do we find in its place?

## 4.6 Steiner's theory of mathematical explanation

Steiner 1978 offers the following theory about the difference between explanatory and non-explanatory proofs:

My proposal is that an explanatory proof makes reference to a characterizing property of an entity or structure mentioned in the theorem, such that from the proof it is evident that the result depends on the property. It must be evident, that is, that if we substitute in the proof a different object of the same domain, the theorem collapses; more, *we should be able to see as we vary the object how the theorem changes in response*. ([127], p. 143; emphasis added.)

Steiner's theory, especially the part I have emphasised, already appears to be an interventionist account. Where he speaks of varying the mathematical objects and seeing how the theorem changes in return, Woodward speaks of varying a causal variable  $X$  and seeing how the variable  $Y$  changes in response. Of course, we would like to know what it means to vary an object; and what the 'characterizing properties' of an entity 'mentioned in the theorem' are. So let us listen to what Steiner says about the example from the previous section.

[The explanatory proof] that the sum of the first  $n$  integers equals  $\frac{n(n+1)}{2}$  proceed[s] from characterizing properties: ... by characterizing the symmetry properties of the sum  $1 + 2 + \dots + n$ . ... By varying the symmetry ... we obtain new results conforming to our scheme. The proof by induction [on the other hand] does not characterize anything mentioned in the theorem. Induction, it is true, characterizes the *set* of all natural numbers; but this *set* is not mentioned in the theorem. ([127], pp. 144-145.)

Steiner accurately captures why the second proof is explanatory: the proof makes us aware of the symmetry of the  $S(n)$ , and how this symmetry is (in a certain sense) responsible for the theorem. What this proof does is to point out a property of the sequence which we can vary to change the result; and although it does not go into explicit detail about the effects of such variation, the competent reader will be able to provide such details herself.

But Steiner's reasons to dismiss the inductive proof as non-explanatory are weak. Claiming that that proof does not explain because the set of all natural numbers is not mentioned in the theorem is a dubious move, not least because a fully formal version of the proof would have started with " $\forall n \in \mathbb{N} : \dots$ ". When written down formally, the set of natural numbers *would* have been mentioned in the proof. So Steiner must be mistaken.

Another weakness in his account is the mysterious notion of a 'characterizing property'. Steiner does not define this notion in his paper, and it

is not clear what we are to make of it. Is it obvious that having a certain symmetry is a characterising property of the sequence  $S(n)$ ? It is certainly a property, but what makes it ‘characterising’? We could also wonder what it means to ‘vary an object’. So although Steiner seems to be on the right track when he gives an interventionist account of the explanatoriness of the second proof, his theory nevertheless misses the strength and the rigour to give a full account of mathematical explanation.<sup>7</sup> We will now see whether a Woodwardian theory can do better.

## 4.7 An interventionist theory of mathematical explanation

In Woodward’s theory, explananda have the form  $Y = y$ , where  $Y$  is a variable and  $y$  is one of its possible values. An explanans consists of a variable  $X$  and its actual value  $x$ , plus a generalisation  $G$  which shows that  $X = x$  implies  $Y = y$  (everything else being equal) and that interventions on  $X$  change  $Y$ . Let us see whether we can cast our example into this mould.

Here is a reconstruction of the proof as a Woodwardian explanation. Let  $X$  be the range of symmetry properties a sequence-scheme of natural numbers

---

<sup>7</sup>Hafner and Mancosu (2005 [36]) have an additional argument against Steiner’s theory: they claim to possess a counterexample. This counterexample is worth discussing here since (if truly a counterexample) it also refutes my Woodwardian theory of mathematical explanation. Unfortunately, it is highly involved, and I cannot repeat all the details here. Therefore, I will limit myself to a few remarks for those who already know the article.

Hafner and Mancosu argue that the arbitrary sequences of positive numbers which feature in Kummer’s convergence test have (because of their very arbitrariness) no characterising properties that could be varied. Therefore, Steiner cannot account for explanations of the validity of Kummer’s convergence test.

I have two remarks to make. The first is that it is simply not true that Kummer’s sequences have no characterising properties. What is true is that no varying of any single number in a sequence would make a difference to its usability in Kummer’s convergence test. But there *is* a property of the sequence which *can* be varied and which *does* make a difference: the property of consisting (after a certain point) of only positive numbers. Sequences that have this property can be used to test convergence, but other sequences, like 1, -1, 1 -1, . . . , cannot. So the answer to “Why can we use 1, 2, 3, 4, . . . in Kummer’s convergence test?” will mention that this sequence belongs to the class of sequences that, after some arbitrary number of elements, consist of only positive numbers. Steiner’s theory can capture this, and so can my generalised interventionist theory.

The second remark is that both Steiner and I are committed to the following: if the question were to arise why *all* sequences share a certain property, the only possible explanations of this would lie in the difference between sequences (as such) and some class of non-sequences. But this seems reasonable enough, and the example of Hafner and Mancosu is not a counterexample to our claim.

(with  $n \in \mathbf{N}$  the only unbound variable in the scheme) can have.  $X$  will include such options as ‘no symmetry’, ‘adding the first to the final term, and the second to the final-but-one, and so forth, always gives the answer  $2n + 4$ ’ and ‘adding the first to the final term, and the second to the final-but-one, and so forth, always gives the answer  $n + 1$ ’. We will call the last option  $x$ . Let  $Y$  range over all theorems of the form ‘The sum of all the terms of the (finite) instantiations of sequence-scheme  $\alpha$  of natural numbers (with  $n \in \mathbf{N}$  the only unbound variable of the scheme) is  $\beta$ ’. Let  $y$  be the particular instantiation ‘The sum of all the terms of the (finite) instantiations of the sequences  $1, 2, 3, \dots, n$  is  $\frac{n(n+1)}{2}$ .’ And let  $G$  be (an appropriate formalisation of) the rule that if pairwise addition of the terms of a sequence always leads to result  $\gamma$ , then the sum of the sequence is the number of terms times  $\frac{\gamma}{2}$ .

$G$  allows us to show that  $X = x$  implies  $Y = y$ , and it also allows us to see that interventions on  $X$  – understood as substituting sequence-schemes with different symmetry properties in the proof – will change the result of the proof. What’s more,  $G$  allows us to see *how* certain interventions will change the result; for instance, what will happen if we take a sequence with the symmetry property that adding the terms pairwise always leads to the result  $2n + 2$ , or  $n! + \ln n$ .  $G$  helps us to see that by constructing sequences with symmetry properties like  $x$ , we will arrive at results like  $y$ . I submit that this is what makes the second proof explanatory.

However, two remarks. First, Steiner makes a mistake if he means to say that proofs *as such* can be explanatory. An explanation is always more than just a proof: it points beyond a proof of  $y$  by showing us that a generalisation  $G$  was used in the proof, and that this generalisation can also be used to prove alternatives to  $y$ . What makes it tempting to speak of an ‘explanatory proof’ in examples such as the one we have been discussing, is that these additional steps are easily seen once the proof is given, so that the intelligent reader will not need them to be pointed out. But strictly speaking, a proof does not explain; it can only suggest or be part of an explanation.

Second, the Woodwardian reconstruction I have just given is not the only possible one. Instead of  $G$ , one could have chosen a more or less general statement, and one might still have a good explanation. This ambiguity was already implied in the previous remark, where we saw that  $G$  had been given only implicitly in the proof. And if you look for them, there are more ambiguities:  $X$  and  $Y$  could be characterised in a slightly different way as well while leaving the proof the same. None of this is a problem for our theory; it merely reminds us how important it is to be clear about the explanandum that a given candidate explanation is supposed to explain, and it shows us that the proofs we discussed did not do this (and were thus not

full explanations).

With these remarks out of the way, we can go on to the more essential task of accounting for the non-explanatory nature of the first proof. Here, the ambiguities we just noted are more malicious, for we *can* reconstruct the proof in such a way that it *does* explain something. Suppose that someone asks the following question: ‘I see that for any  $n$  you can prove that  $S(n) = \frac{n(n+1)}{2}$ . But what allows you to conclude that  $\forall n : S(n) = \frac{n(n+1)}{2}$ ?’<sup>8</sup> In that case, the first proof *would* be an explanation, at least once the implicit statements that complete induction characterises the natural numbers and that changing this property of the domain of our theorem would make the theorem invalid had been made explicit. Of course the first proof does *not* explain  $Y = y$  (as defined in the previous paragraphs), since varying whether or not the principle of induction holds does not make a difference within  $Y$ . So the first proof can be used in an explanation; it just cannot be used to explain what most naturally puzzles us when we see the theorem, namely, that  $Y = y$ . (Using a Woodwardian approach has thus already allowed us to see beyond our first intuitions.)

All of this suggests that the formal structure of Woodward’s theory of causal explanation can be transferred to the domain of mathematics, but we have not yet said anything about the notions of ‘intervention’ and ‘cause’. Before we can speak of an interventionist theory of mathematical explanation, this defect must be remedied.

We will use the notion of ‘intervention’ as a general term not limited to causal interventions. The notion of ‘cause’, also prominent in Woodward’s theory, will be replaced by the more general notion of ‘ground’. It is to be understood that a cause is a kind of ground; that a direct cause is a kind of direct ground; and so on. Like Woodward, we will use ‘intervention’ in order to (semi-circularly) define ‘ground’.

But why do we even need a notion of ‘ground’ in a theory of mathematical explanation? For the same reason that we need the notion of ‘cause’ in a theory about causal explanations: to generate asymmetries. The notion of cause helps us to show why the length of the shadow is explained by but does not explain the length of the flagpole. The more general notion of ground will have to help us show why the fact that the natural numbers are characterised by induction (let us call this  $A$ ) helps us to explain the truth of  $\forall n : S(n) = \frac{n(n+1)}{2}$  (and let us call this  $B$ ), but not the other way around.

We now formally define the notions of direct ground, indirect ground, intervention variable and intervention in our interventionist theory of math-

---

<sup>8</sup>This is not a trivial question. The proof would, for instance, not be possible in Peano arithmetic without the axiom of induction.

ematical explanation by stipulating that they satisfy all Woodward's definitions of direct cause, indirect cause, intervention variable and intervention. We also stipulate that the variables used in these definitions range not over type-level events, but over mathematical objects<sup>9</sup> and their properties.

So in part,  $A$  is a (direct or indirect) ground of  $B$  because changing  $A$  will change  $B$ . But what does it mean to change whether or not  $A$ ? And what does it mean to say that changing  $A$  will change  $B$ ? (What it does *not* mean is that certain alternatives to  $A$  are inconsistent with  $B$ ; for being inconsistent with is a symmetrical relation, and grounding emphatically is not.) Answering these questions is essential. We need to look at the practice of mathematical reasoning and identify the kinds of intervention that actually occur in it.

1. We often start doing mathematics by identifying the mathematical objects about which we wish to learn more. We choose to investigate the natural numbers, say, or the geometry of Euclidean plane figures. Depending on one's philosophy of mathematics, one will describe this choice differently: as choosing formal axioms, as choosing rules of construction, as choosing properties by which to identify the mathematical objects in a Platonic realm. Whatever the description, choosing these axioms, rules or properties is the most basic intervention in mathematical research, and these axioms, rules or properties are grounds for everything that follows from them. Example: Choosing the axiom of complete induction is part of choosing to investigate the natural numbers. Therefore, this axiom can be a ground for the fact that for all natural numbers,  $S(n) = \frac{n(n+1)}{2}$ , but not the other way around.
2. Within a mathematical realm thus delimited, we construct (identify) mathematical entities and prove theorems about their properties. The choices we make during the construction are interventions; and the properties explicitly associated with these interventions are grounds for the properties that follow from them. For example, let  $P$  be the property that a sequence of numbers has if and only if, starting from the beginning and the end, pairwise addition of the terms of the sequence always gives  $1+n$ , where  $n$  is the number of terms in the sequence.  $P$  is a 'constructive' property in the sense that it is immediately clear, without giving any further proofs or constructions, how we can construct sequences with this property.<sup>10</sup> The property  $P$  can then be proved

---

<sup>9</sup>If one believes that the type-token distinction makes sense for mathematical objects, we mean types here.

<sup>10</sup>The notion of 'immediate clarity' involves a certain subjectivity, or at least implies mathematical non-omniscience. See page 186 for further discussion of this point.

to imply the property  $S$ : that the sum of the terms of a sequence is  $\frac{n(n+1)}{2}$ . Hence, constructing an entity with  $P$  is a way to construct an entity with  $S$ , and therefore,  $P$  is a ground of  $S$ .<sup>11</sup>

These two material characterisations of what intervention is in mathematics, when combined with the formal characterisation already laid out, will suffice to start analysing mathematical explanations in terms of interventions and grounds. Whether Woodward's theory thus applied to mathematics will lead to correct results, identifying all and only 'explanatory proofs' as explanations, cannot be argued further in this place; only a multiplication of successful instances and the absence of unsuccessful ones will produce conviction. Instead, we will look at a second kind of non-causal explanation.

## 4.8 Law explanations

With the term 'law explanations', I wish to designate explanations of laws of nature by other laws of nature. The classic example is the explanation of Kepler's laws by deriving them from the Newtonian laws of motion. This kind of explanation is not causal: laws of nature are not events, and cannot be each other's causes or effects. Law explanations are thus a second case of explanations which cannot be captured by Woodward's original theory, but should be captured by our generalised version of it.

Our general strategy will be the one we developed in the previous section: we use the formal structure of Woodward's theory in order to speak about interventions on and relations of grounding between laws of nature; then, we give a partial interpretation of these notions which is strong enough to get the explanatory discourse going. Preferably, we then check the outcome against many examples – but for reasons of space I will have to focus on the single example of Kepler's and Newton's laws. In this section, though, I will forgo spelling out the formal machinery. Instead, I will focus on the need to find a notion of intervention that is applicable to laws of nature. This notion must generate the asymmetry that Newton's laws explain their consequences, but that those consequences do not in turn explain Newton's laws.

We cannot think of interventions on laws of nature in the same way we think of interventions on states of affairs. One might even believe that it is a

---

<sup>11</sup>We are not just giving a new name to the notion of logical implication. Grounds need not imply their consequences, and consequences may imply their grounds. A shadow implies a source of light, but a source of light does not imply a shadow; and yet the source of light is a ground of the shadow, because we can intervene on sources of light to create shadows, but we cannot intervene on shadows to create sources of light.

necessary (though not a sufficient) condition for lawhood that the regularity in question can *not* be changed through the physical interventions that we use to intervene on states of affairs; they are beyond our control. How then are we to conceive of interventions on laws of nature? The answer will partly depend on one's theory of lawhood, but for our purposes, the metaphysical questions can be skirted. All we need in order to show that our generalised interventionist theory can handle law explanations, is a criterion for the truth of statements describing the results of interventions on laws of nature. If we know whether it is true or false that "An intervention changing law  $X$  to law  $X'$  has the result that  $Y$  changes to  $Y'$ ", for any  $X$ ,  $X'$ ,  $Y$  and  $Y'$ , we will be able to use the interventionist theory of explanation to assess law explanations. It is irrelevant for these purposes that we are not in fact in a position to change any law whatsoever. Now I will suggest that we possess strong intuitions about the truth of counterfactual statements involving laws; and that we can use these intuitions to judge the truth of statements about the results of interventions on laws.

This strategy will work only if there is a strong link between intervention and counterfactuals – otherwise, linking up laws and counterfactuals will not result in an *interventionist* theory of law explanations. This strong link exists, and James Woodward has already done the hard work of spelling it out. Since I lack the space to repeat his discussion here, I refer the reader to his book, especially pp. 279-285.<sup>12</sup>

The link I want to forge between counterfactuals and interventions on laws of nature is captured by the following meaning postulate: "An intervention changing law  $X$  to law  $X'$  has the result that  $Y$  changes to  $Y'$ ." means "If, contrary to fact,  $X'$  rather than  $X$  were to be a law of nature, then  $Y'$  rather than  $Y$  would hold." Of course, this clears things up only if the truth or falsity of the second kind of statement is easier to ascertain than that of the first kind; and this will only be the case if we have solid intuitions about the truth values of this second kind of statement. Luckily, we do. We may not have a satisfactory *theory* of counterfactuals, but that is an additional and separate problem; our intuitions about the truth values of individual counterfactuals

---

<sup>12</sup>Perhaps the main reason that we believe Newton's laws to be more fundamental than Kepler's laws is that the latter are a specialised version of the former, which we get by putting in certain boundary conditions. Thus, to explain the explanatory asymmetry in terms of counterfactuals may seem a terribly roundabout way of arriving at the point. Yet this is not the case; for we need the notion of 'counterfactuals' in order to arrive at the notion of 'intervention'; neither 'boundary condition' nor 'specialised version' will take us there. It is because the difference between Newton's laws and Kepler's laws can be stated in terms of counterfactuals that it can be stated in terms of interventions, and thus subsumed in our interventionist theory of explanation.

are often quite clear. Let us look at the Newton-Kepler example.

In order to get the right conclusions about explanation, our theory would have to say that under many interventions on Newton's laws, Kepler's laws would change; but that, on the other hand, Newton's laws remain invariant under all interventions on Kepler's laws. Well, how do the counterfactuals come out in this case? Suppose, counterfactually, that Newton's law of gravity related gravitational force not to the inverse square of the distance, but to the inverse cube of the distance. What kind of orbit would the planets then make around the sun? We'd have to solve a non-trivial differential equation, but the answer is surely that the planets would not be moving in ellipses.

On the other hand, suppose that contrary to fact the planets were moving around the sun in equilateral triangles. What is the closest world in which this counterfactual is true? Let me present four candidates. In world *A*, the laws of motion and gravity in the Universe are such that all objects move in a straight line until some condition is met, at which point they make a 60 degree turn and continue in the new direction until the condition is met again, and so on. In world *B*, in addition to our force of gravity, there is another force which influences the planets (but not much else); these forces work together to determine a triangular orbit. In world *C*, huge jet engines have been installed on the planets to change their course into triangles. In world *D*, the planets do not go around in ellipses because they have been tethered together with large chains.

We may not have clear intuitions about which one of these four scenarios is closest to our actual world. But I do think we have clear intuitions that of the four scenarios, *A* is farther from our actual world than *B*, *C* and *D*: only in world *A* are the laws of nature substantially different, and only in world *A* will most events in space-time be different (for it would be a miracle if things worked out more or less the same while the fundamental laws of motion were very different). That is all we need. We need to show that we do not judge that a violation of Kepler's laws counterfactually implies a violation of Newton's laws; and that is just to show that we judge worlds like *A* to be farther from our actual world than worlds like *B*, *C* and *D*.

It may be objected to my example that I am relying excessively on intuitions without making clear what the truth conditions of counterfactuals are, and how they relate to laws. But I don't need to do that. All I *need* to show in this chapter is that our intuitive judgments about which laws explain which line up with our intuitive judgments about interventions on laws. What I do not need to show is that these intuitions are right. It is of course possible that a successful theory of counterfactuals is thought up which convinces us that the intuitions I used here are wrong; in that case, we will have to reassess the interventionist theory (and see, particularly, if this

theory of counterfactuals also changes our ideas about which laws explain which). But it is unreasonable to ask us to do so in advance, before such a theory has been accepted.

We will now return to the ambition laid out in the earlier part of this chapter: to give a full generalisation of Woodward's theory that is applicable to all explanations, whether causal, mathematical, or of whatever other kind.

## 4.9 A general interventionist theory of explanation

What we have seen in the previous three sections is that a theory structurally equivalent to Woodward's interventionist theory of causal explanation can also be used to explain mathematical explanation and explanations by laws of nature. We will now write down a formal generalisation of his theory, the ambition of which is to cover all explanations. I will not attempt to give a material characterisation of the notion of intervention in each and every explanatory domain; but I will give a few constraints on such material characterisations and assert that if the formal rules and the material constraints are satisfied by any characterisation of intervention, then the characterisation is valid and it applies to a domain of explanation.

First the formal part; most of it comes straight from Woodward. Let  $\mathcal{F}$  be a set of variables of appropriate generality.<sup>13</sup> Let  $\mathcal{G}$  be a set of generalisations which state that certain values of certain of these variables imply certain values of certain others. Let  $\mathcal{F}^*$  be the set of all the members of  $\mathcal{F}$  which appear on the left-hand side of the implication sign in members of  $\mathcal{G}$ . ( $\mathcal{F}^*$  contains all the members of  $\mathcal{F}$  which are candidates for being grounds.)

Now, we define grounds:

A necessary and sufficient condition for  $X \in \mathcal{F}$  to be a **direct ground** of  $Y \in \mathcal{F}$  with respect to a variable set  $\mathbf{V} \subset \mathcal{F}$  is that there be a possible intervention on  $X$  with respect to  $Y$  that will change  $Y$  or the probability distribution of  $Y$  when one holds fixed at some value all other variables  $Z_i$  in  $\mathbf{V}$ . A necessary and

---

<sup>13</sup>Woodward formulates his theory in terms of type-level events because causal generalisations are (almost always) made on that level. I formulated the mathematical version of the theory in terms of the properties of mathematical entities, because the mathematical generalisations used to give explanations are made on that level; even so, a formalist and a Platonist would differ in their semantic analysis of such explanations, and hence would differ on what a mathematical property is. In general, the variables in  $\mathcal{F}$  must be of the sort that occur in the generalisations used to theorise about the explanatory domain.

sufficient condition for  $X \in \mathcal{F}$  to be a **contributing ground** of  $Y \in \mathcal{F}$  with respect to a variable set  $\mathbf{V} \subset \mathcal{F}$  is that (i) there be a directed path from  $X$  to  $Y$  such that each link in this path is a directed causal relationship; that is, a set of variables  $Z_1 \dots Z_n$  such that  $X$  is a direct ground of  $Z_1$ , which is in turn a direct ground of  $Z_2$ , which is a direct ground of  $\dots Z_n$ , which is a direct ground of  $Y$ , and that (ii) there be some intervention on  $X$  with respect to  $Y$  that will change  $Y$  when all other variables in  $\mathbf{V}$  that are not on this path are fixed at some value.

Next, we give the formal characterisation of intervention variables

$I \in \mathcal{F}$  is an **intervention variable** for  $X \in \mathcal{F}$  with respect to  $Y \in \mathcal{F}$  if and only if  $I$  meets the following conditions:

- I1.  $I$  grounds (is a ground of)  $X$ .
- I2.  $I$  acts as a switch for all the other variables that ground  $X$ . That is, certain values of  $I$  are such that when  $I$  attains those values,  $X$  ceases to depend on the values of other variables that ground  $X$  and instead depends only on the value taken by  $I$ .
- I3. Any chain of direct grounds from  $I$  to  $Y$  goes through  $X$ .
- I4.  $I$  is (statistically) independent of any variable  $Z$  that grounds  $Y$  and that is on a directed path that does not go through  $X$ .

and of intervention

$I$ 's assuming some value  $I = z_i$ , is an **intervention** on  $X$  with respect to  $Y$  if and only if  $I$  is an intervention variable for  $X$  with respect to  $Y$  and  $I = z_i$  is an actual ground of the value taken by  $X$ .

Given these preliminaries, the formal notion of explanation becomes:

Suppose that  $M$  is an explanandum consisting in the statement that some variable  $Y \in \mathcal{F}$  takes the particular value  $y$ . Then an **explanans**  $E$  for  $M$  will consist of (a) a generalization  $G \in \mathcal{G}$  relating changes in the value(s) of a variable  $X$  (where  $X$  may itself be a vector or n-tuple of variables  $X_i \in \mathcal{F}$ ) and changes in  $Y$ , and (b) a statement that the variable  $X$  takes the particular value

$x$ . A necessary and sufficient condition for  $E$  to be (minimally) **explanatory** with respect to  $M$  is that (i)  $E$  and  $M$  be true or approximately so; (ii) according to  $G$ ,  $Y$  takes the value  $y$  under an intervention in which  $X$  takes the value  $x$ ; (iii) there is some intervention that changes the value of  $X$  from  $x$  to  $x'$  where  $x \neq x'$ , with  $G$  correctly describing the value  $y'$  that  $Y$  would assume under this intervention, where  $y' \neq y$ .

We then turn to the material constraints on intervention. The interpretations of intervention in a certain domain must (i) be consistent with all the claims of the formal theory given above, (ii) be consistent with our counterfactual intuitions (or our accepted theories of counterfactuals) about changes on members of  $\mathcal{F}^*$  and the results of these changes, (iii) be such that for every member of  $\mathcal{F}^*$  it makes sense (that is, it is not meaningless or nonsensical) to speak of intervening on that variable, and (iv) not be a gerrymandered notion. If no such interpretation of intervention can be found, explanations are not possible within the explanatory domain specified by  $\mathcal{F}^*$  and  $\mathcal{G}$ .

This is the generalised interventionist theory of explanation. Both Woodward's theory and the theories of mathematical explanation and explanation by laws of nature I have given are instantiations of it, as the reader can verify. Will it also work for other forms of explanation? Only attempts to actually apply it to them will show.

## 4.10 Conclusion

In this chapter I have discussed Woodward's interventionist theory of causal explanation, adapted it to mathematical explanation and explanations by laws of nature, and shown how it might be generalised to cover all explanations. These efforts fall far short of actually demonstrating this generalised theory to be correct; but they open the possibility of further research in this direction. If the generalised theory expounded here will be found successful, we will have made an important step forward in understanding explanation. If it will be found lacking, we will at least gain insight into what makes causal explanation special.

In the context of this thesis, the generalised interventionist theory should be seen as a preliminary version of the determination theory expounded in chapter 6. The notion of intervention used in that chapter will be exactly as defined above. However, the determination theory will drop the idea that explanations must use generalisations; rather than demand that there be some intervention turning  $x$  to  $x'$  that changes the value of  $Y$  from  $y$  to  $y'$ ,

it will demand only that this intervention reduces the probability of  $Y$  being  $y$  to less than 1; rather than demand that there be some such intervention for *some*  $x' \neq x$ , it will demand that some such intervention exists for *all*  $x' \neq x$  (but it will simultaneously introduce a formalism that allows us to coarse-grain variables); and it will more clearly contain and embrace the idea that all explanations can be given in the form of deductions.

