Stockholm University

Taking Control

The role of manipulation in theories of causation

Henning Strandin

Doctoral Thesis in Theoretical Philosophy at Stockholm University, Sweden 2020

Taking Control

The role of manipulation in theories of causation

Henning Strandin

Academic dissertation for the Degree of Doctor of Philosophy in Theoretical Philosophy at Stockholm University to be publicly defended on Thursday 9 January 2020 at 13.00 in Hörsal 7, hus D, Frescativägen 10 D.

Abstract

Causation has always been a philosophically controversial subject matter. While David Hume's empiricist account of causation has been the dominant influence in analytic philosophy and science during modern times, a minority view has instead connected causation essentially to agency and manipulation. A related approach has for the first time gained widespread popularity in recent years, due to new powerful theories of causal inference in science that are based in a technical notion of *intervention*, and James Woodward's closely connected interventionist theory of causation in philosophy. This monograph assesses five manipulationist or interventionist theories of causation, viewed as theories that purport to tell us what causation is by providing us with the meaning of causal claims. It is shown that they cannot do this, as the conditions on causation that they impose are too weak, mainly due to ineliminable circularities in their definitions of causal terms. It is then argued that a subset of Woodward's theory can nevertheless contribute crucially to an explanation of the unique role that manipulation has in our acquisition of causal knowledge. This explanation differs from the common regularist explanation of the epistemic utility of manipulation and experiment, and it is taken to confirm several important manipulationist intuitions. However, the success of the explanation depends on (this subset of) interventionism not itself being understood as a theory of causation, but as a theory of intervention.

Keywords: *philosophy of science, causation, causal models, manipulationism, interventionism.*

Stockholm 2020
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-176255

ISBN 978-91-7797-913-5
ISBN 978-91-7797-914-2

Department of Philosophy

Stockholm University, 106 91 Stockholm

TAKING CONTROL

Henning Strandin

Stockholm
University

Taking Control

The role of manipulation in theories of causation

Henning Strandin

ISBN print 978-91-7797-913-5
ISBN PDF 978-91-7797-914-2

Printed in Sweden by Universitetsservice US-AB, Stockholm 2019

For patient and generous
people everywhere.

Acknowledgements

While my actions can plausibly (or at least conventionally) be judged the salient cause of this monograph, its coming about has essentially depended also on input from a number of other people.

The many detailed comments and suggestions offered over the years by my supervisors Peter Pagin, Richard Dawid, and Paul Needham have of course been indispensable. The internal reviewers Sebastian Lutz and Erik Angner provided important feedback in the finishing stage.

Having the opportunity to regularly present ideas at our local seminars has been instrumental, and my thanks go out to the organizers and participants of the PhD Seminar, the Joint Stockholm/Uppsala Seminar in the Philosophy of Science, the CLLAM Seminar, and the Higher Seminar in Theoretical Philosophy.

I've benefited greatly from discussions with my colleagues and friends Eric Johannesson, Sara Packalén, Sama Agahi, Anders Lundstedt, Mattias Högström, Björn Brunnander, and many others. The open and friendly atmosphere at the Stockholm University philosophy department has made this enjoyable work in general. Completing a PhD thesis also depends on the generosity and patience of family and friends, and I have had an abundance of this.

Finally, the production of this text was made significantly more convenient by the free use of the Open Source LaTeX editor LyX, and I owe its developers a thanks.

Contents

CHAPTER 1

Introduction

1.1. General introduction to the monograph

Several years ago, during a seminar consisting of PhD students, and some senior faculty including my supervisors, I was talking about Armstrong's defense of laws as relations between universals. Armstrong famously claimed that laws of the kind he advocated could, while laws understood as regularities could not, *explain* certain regularities we observe, as well as probabilistic and counterfactual dependencies we associate with some regularities (1983). I injected at one point, not as an objection to Armstrong I think, but as a matter of clarification, that we don't tend to expect *all* observed regularities to have such an explanation, but only those we have *tested* somehow, as with an experiment. That is, some regularities tend to persist despite our best efforts to break them, others haven't been tested in this way. Only the former would strictly speaking require an explanation in terms of a law. I, naively, expected this to be a straightforward and uncontroversial point. But several participants at the seminar spontaneously disagreed with the sort of role I gave to *manipulations* and *experiments* in this context. The alternative that they found more plausible is just that when we perform experiments we *see different things*, as compared to when we observe passively. Manipulation as such, then, plays no essential part in the difference between the two situations. I understand now that my original intuition was anything but straightforward and uncontroversial. As I was already writing in the philosophy of causation, this unexpected interchange led me to reformulate the question for my dissertation: *what is the role of manipulation in theories of causation?* Thus, I'm largely indebted to those seminar participants for whatever substance this monograph may have.

As I began to examine manipulationist theories of what causation *is*, as well as some scientific methods of causal inference, I had to consider a different role for manipulations in theories of causation. Namely, that manipulation, or the possibility of manipulation, or just agency in general, is an essential part of the *nature of causation*. Or, closely related, that the possibility of manipulation is an ineliminable part of the *content* of *causal concepts* or *causal claims*. This issue would become the focus of most of the monograph, while I retained also the question of how to precisely understand the difference—if there is one—that manipulation could make *epistemically* and *methodologically*, in our relationship to causation.

While the older manipulationist theories of causation, for example by Collingwood and von Wright, were never embraced by the philosophical mainstream, my project took on additional relevance, it seemed to me, from the fact that James Woodward's more recent *interventionist* theory of causation has attracted broad interest, while also connecting more explicitly than earlier incarnations to specific scientific methods.

Partly due to the nature of my original question, and partly, I'm sure, just due to the sort of training I've received, my focus has largely been on the *logic* of these theories. That is to say, what is the logical form of the conditions they present, and what sorts of consequences do they have, especially in relation to classical problems in the philosophy of causation, such as the identification of the direction of a causal connection or the exclusion of common causes and "accidental" associations. And how do these condition eventually succeed as a *definition* of causation, if they do? This has also constrained, to some extent, the literature that I've considered most relevant to my own inquiry.

The rest of this chapter provides a brief overview of the features of manipulationism that are of main interest to our discussion, and of how this kind of theory relates to more mainstream concerns in the philosophy of causation. In chapters 2 to 7 I assess the most well-known manipulationist treatments of causation and some criticisms of them, with an interlude in chapter 6 where I outline *causal models*, to provide some additional context for Woodward's theory. In chapter 8 I give my own main arguments. I argue that the most important objection to earlier manipulationist theories of causation does not concern their anthropocentrism (or -morphism), or their inherent circularity, but the fact that they are inadequate as ordinarily stated because they don't provide a sufficient condition for causation.

The main part of that chapter, however, contains my own assessment of Woodward's interventionist account. I argue in that chapter that Woodward's analysis is viciously circular in a well-defined way, and that it therefore can't provide an explicit definition of causation, or a plausible "connective analysis" of the concept. It is implied, as well, that if the earlier manipulationist theories are made adequate, by their conditions for causation being made sufficient, they suffer the same fate as Woodward's account in this respect. (I.e., they become viciously circular.) I moreover show that Woodward's theory also can't succeed as an inductive or an implicit definition of causation.

Despite this, Woodward's theory implies a certain sufficient condition for causation that I argue is true of causation as it is in the world, and I show that this part of the theory can be retained, and the remainder either jettisoned or given a very different interpretation from the first part. The resulting theory can't deliver an understanding of what causation is, or of what causal claims mean, but it can underwrite sound causal inference rules and also help us understand the *epistemic* and *methodological* role of manipulation and experiment in our causal knowledge acquisition. I propose, then, that a certain modified version of Woodward's theory is plausibly true of causation, but more usefully viewed as a *theory for causal inference under intervention* and thereby as a *theory of intervention*. Toward the end the chapter, I substantiate the

claim that this theory can aid our understanding of how we can acquire causal knowledge when we manipulate things, by providing a sketch of such an explanation. Such an explanation, if true, accomplishes two important things. It to a considerable extent validates what I take to be the intuitions driving the manipulationist approach in the philosophy of causation. And it shows "cause" to be an *empirical notion* that, from an empiricist perspective, therefore doesn't require a definition or analysis to be meaningful.

I end up, then, rejecting the role given to manipulations in manipulationist accounts of the nature of causation, and the meaning of causal claims. But, equally important, I also end up rejecting a certain traditional—largely regularist—view of the *epistemic* role of manipulation and experiment, in science and everyday life. That view is just that whether we personally manipulate something makes no difference—all that matters is what we *observe* as a result. I argue that this is wrong, and on this point I therefore agree with the manipulationist intuitions. Hopefully, I will have managed, by the end of this monograph, to also explain what I had in mind at that seminar, when claiming that some regularities and not others are of particular interest, because they have been subject to a special kind of test, involving manipulation or experimentation.

The final chapter summarizes my conclusions and suggests some questions, both empirical and philosophical, that seem good candidates for future research.

1.2. Manipulationism: general motivations

Manipulation theories of causation are traditionally characterized by essentially associating causation with manipulations and agency. A recent development is a sort of theory that associates causation with *interventions* rather than manipulations. "Intervention" is a technical term in this context, and refers in Woodward's theory to an event that has certain causal relations, and lacks certain causal relations, relative to other parts of a causal system. Thus, a manipulation viewed as a voluntary action by an agent can fail to satisfy the conditions on an intervention, and an intervention in turn can be a natural event that doesn't involve human or other agents. The concepts overlap, when a manipulation is an intervention, but neither concept contains the other. However, because interventions are a sort of generalization of manipulations, stemming mainly from the notion of an idealized experiment, and because Woodward in particular calls his interventionist theory a manipulationist theory (2003), I will use "manipulationism" as the more general label for this class of theories.

The first modern manipulation theories appear one or more decades before more familiar theories in the philosophy of causation, such as J. L. Mackie's INUS-analysis (Mackie 1965), David Lewis's counterfactual analysis (Lewis 1987b), and Wesley Salmon's mark transfer or process theory (Salmon 1984). The rivals at the time were thus rather Hume-inspired and Millian-style regularity analyses of causation, such as that suggested by the covering-law model of causal explanation (Hempel and Oppenheim 1948).

I don't believe that manipulationist theories of causation were developed mainly as solutions to issues identified in connection with empiricist and regularist treatments of the topic. The manipulationist motivations are quite different from those driving empiricist and regularist accounts. However, as certain problems, mainly associated with these more mainstream approaches, began to loom large in the philosophy of causation in general, responding to them in a satisfactory way became an adequacy condition also for manipulationist proposals. And it has appeared to some authors that a manipulationist analysis can give a better answer to some of these hard questions than a regularist account could.

Broadly speaking, two kinds of reasons can then be found to motivate manipulation theories. Firstly, establishing the connection between causation and agency can be seen as a goal in its own right; this can appear to these writers to be the correct understanding of the concept of causation, based on for example intuitions about everyday and scientific causal judgments, scientific practices (this motivation is prominent especially in recent intervention theories), or the etymology of causal terms. The important intuitions concerning causation in this case are then radically different from those that for example Hume and Mill, and later regularists, relied on. The most important difference is that the Hume-Mill theory is "passivist" in nature, in the sense that it treats believers in causation as passive observers of regularly co-occurring event types, while it is essential to manipulationists that believers in causation are also agents, interacting with the world through their voluntary actions. Although explicit manipulation theories of causation were first proposed as recently as the first half of the 20th century, the idea that causation is essentially connected to agency can be traced back to antiquity, in Aristotle's teleological explanations of natural phenomena, and found later in Thomas Reid's suggestion that causal power implies the power to *choose to do otherwise* (Reid 2010, Essay I, ch. V).

Secondly, manipulation theories have been taken to resolve certain problems in the philosophy of causation. This is the topic of section 1.4, below. These two motivations can be given different weight in different theories. But in each case, the suggestion that the manipulationist theory can resolve well-known theoretical problems is taken to support the view that the manipulationist treatment of causation is the correct one.

There is a third, important, way of associating manipulations with causation. It has seemed that manipulations often play an important role when we acquire new causal knowledge, especially in the context of scientific experiments. However, as this role is epistemic and not ontological or metaphysical, it is compatible with different theories about how exactly this works, and about what causation *is*, specifically. Still, the epistemic importance of manipulations is constantly at the surface of manipulation theories of causation.

1.3. Defining manipulationism

First a typographical note: I will denote particular events by lower-case typewriter letters (`a`, `b`, ...) and event types by upper-case typewriter letters (`A`, `B`, ...). In what follows, I have sometimes changed the typography of quotes

where it is clear that what is talked about are particular events, or event types, in accordance with this, for consistency.

A manipulation theory about singular causation will typically say that `a` is a cause of `b` *if and only if* it is the case that, were a *manipulation* to bring about/prevent `a`, then `b` would occur/fail to occur. There are, however, many particular causal situations in which the purported cause apparently *cannot* be manipulated, such as those involving the gravitational effects of planet-sized objects, or where the putative cause is a fundamentally stochastic event. A choice must therefore be made as to how the implication on the right-hand-side of the equivalence is to be understood for such cases. Here are four options:

(1) A theoretical possibility is to interpret the implication materially, in which case the condition is trivially satisfied for all unmanipulable events. This is clearly not what is sought.
(2) Another alternative is to say that the claim as to what would be the case if the purported cause was manipulated is *meaningless* when it in fact cannot be manipulated.
(3) Yet another option is to deny this appearance of non-manipulability, and insist that every cause *is* manipulable, in the sense required of the theory.
(4) Lastly, there is the option of claiming that it is the relevant *type* of the particular cause that has some manipulable (or even actually manipulated) instances, and that this is the right condition.

This problem looks different in intervention theories, relative to the earlier manipulation theories, as the corresponding condition does not involve manipulations by agents, and thus doesn't engage with the issue of manipulability. But there is a parallel question about the *intervenability* of causes, that sometimes may relate to the requirement that causal systems are *modular*, in a particular sense. We will return to this question in the chapter about James Woodward's theory.

That a theory of causation implies a *sufficient* condition for causation in terms of what happens under a manipulation is not a very good indicator of it being a manipulationist theory, specifically. This is because, under the right understanding of what a manipulation is, causal theorists of many different stripes may agree with such a sufficient condition for causation. I'll therefore suggest the following definition of causal manipulationism, for the context of the coming discussion: a manipulation theory of causation contains as a *necessary* condition for some particular event to be a cause, that it has a certain kind of relation to something that can be manipulated. More specifically, a cause is at least an event of a *type* that can be manipulated, or it is perhaps *composed* of events of such types. This ensures that manipulation (or intervention) plays the sort of essential role in the theory's account of causation, or of the meaning of causal claims, that makes it appropriate to identify it as a special sort of theory of causation, in this respect. Thus, alternatives 3 and 4 in the above list occur in some form in the theories we will consider. (Alternative 2 shows up as a possible interpretation of some, but not all, things Woodward says in

connection with his interventionist treatment.) We can then use the following definition of a manipulation theory of causation:

> **MT:** A theory of causation is a manipulation theory *if and only if* it implies a condition for singular causation of the following kind:
>
> For particular events `a` and `b`, `a` is a cause of `b` *only if* `a` is an event of a type that has instances that are manipulable/intervenable (with respect to `b`), or `a` is composed of such events.

The implied condition is clearly satisfied in cases where the very event that took place could have been changed or prevented, or where a particular possible event thought to be a possible cause, that did not actually occur, could have been brought about through a manipulation or intervention. In such cases, the relevant event type can be very narrow. (The class may just be the singleton of the actual event.) In other cases, the event type will have to be broadened before we find something manipulable, and different manipulation theories differ in how they describe the relevant types. Event types can here not be allowed to be constructed in any arbitrary and disjunctive way, as that could make the manipulability condition vacuous. They must be *natural* in some respect. If an analysis is intended, events can also not be classified on their *causal* properties, on pain of circularity.

There are many examples of writers who recognize the importance of manipulation or intervention in their accounts of causation, but that do not propose an overtly manipulationist theory of causation. One example is the proposal that Peter Spirtes, Clark Glymour, and Richard Scheines make in *Causation, Prediction, and Search* (Spirtes et al. 2000). This is one of the canonical works on interventionist causal inference theory, but they explicitly do not impose the necessary condition on causes stated in **MT**, but leave the question as to how causation is to be defined, if it is to be defined at all, open (Spirtes et al. 2000, p. 19; see also Glymour 2004, p. 788-789). While Judea Pearl gives a *formal* definition of "causal effect" in terms of an intervention in *Causality*, it's clear especially from more recent writings that he is not proposing that causation, or the content of causal concepts, generally depends on possible intervention (Pearl 2009, 2018). Another writer who has emphasized the importance of manipulations to our understanding of causation, but rejected any attempt to *define* causation in manipulation terms, is Nancy Cartwright (e.g., Cartwright 1983, p. 22). D. H. Mellor endorses causes as *means to an end* in *The Facts of Causation*, where he claims that, among the different connotations of "cause," the means-to-an-end connotation is "the very core of the concept" (Mellor 1995, ch. 7). However, Mellor argues that the means-end relation is grounded in (or perhaps just co-extensional with) the same relation of probabilistic dependence as the other useful connotations of "cause," and also that, as an event being an end for someone depends on it having some positive *value* for that person, not every effect is an end. (While Mellor mentions Gasking's theory in this context, his discussion of means and ends specifically is in fact more reminiscent of Collingwood's treatment, as we shall see below.) Thus, as I understand Mellor's theory, **MT** is not implied, and Mellor's is therefore not a manipulationist account of causation, as we define it here.

1.4. Manipulationism and the old problems

Here I will describe in very general terms how manipulation theories have been taken by their authors to resolve some well-known problems in the theories of causation. Not all manipulation theories adopt every one of these approaches. There is especially a tendency to emphasize either the first problem below, in earlier theories that are not put forth mainly in terms of logical conditions, or the later problems, in more logically or mathematically oriented, more recent theories. Note again, that one problem manipulation theorists are likely to have with other theories of causation is just that those other theories do not conceptually associate causation with agency. This essential aspect of the theory is not necessarily of mere instrumental value in solving other problems.

1.4.1. Explaining the impression of necessary connections in nature. Hume gave what would be the standard empiricist argument against the presence of *necessary connections* between natural events. Empiricism implies that the content of every idea is given by experience, either as produced by the senses, or by reflection on ideas formerly produced by the senses. In neither of these sources could Hume find a necessary connection between those events that we believe to be causally connected, or between their ideas. Thus, he proposed that the necessary connection, that we nevertheless as a matter of fact associate with these events, is an impression produced by the mind itself, when we repeatedly experience events of these types occurring in conjunction (Hume 1888, 1.3.14). All that *objectively exists* in these cases, then, are the regular spatiotemporal conjunctions of particular events of the types in question, and there is no objective sense in which the conjunction is *necessary*, and no sense in which one event can *force* or *compel* another event to take place.

Manipulationists who are sympathetic to this latter view, tend nevertheless to reject Hume's explanation of *how* our impression of necessitation in causation arises, that is, from our experience of regular co-occurrence. Rather, they trace this notion of necessity to the human experience of being rationally compelled to *act* in a certain way, to compel others to act, or to the sensation of being the agent behind certain changes in the world, that are consequences of voluntary actions. This idea goes back at least to Reid 2010, and is most explicitly articulated in this text by R. G. Collingwood's theory. It can be seen in some form also in the theories of George Henrik von Wright, and Peter Menzies and Huw Price, although to what extent this "projected" feature of causation is nevertheless a *real* feature of the world differs between the theories.

1.4.2. Saving the meaningfulness of the notion of cause. John Stuart Mill reduced laws, including causal laws, to regularities among particulars, as required by Humean-style empiricism. Nevertheless, the purpose of Mill's inductive method—and hence of science on his account—is to identify causes of phenomena, in the form of regular antecedents to these. Mill, then, in the Humean spirit, denies the existence of any sort of natural necessity relating cause to effect, but insists all the same that the notion of cause is "the root of the whole theory of Induction" and that "the universality of the law of causation

consists in this, that every consequent is connected [...] with some particular antecedent, or set of antecedents" (Mill 1882, book III, ch. V).

In the *Critique of Pure Reason*, Kant had furthermore proposed, in the "Second Analogy," that "All alterations occur in accordance with the law of the connection of cause and effect" (Kant 1998, p. B233). This "principle of causality" was regarded by Kant as a *synthetic a priori*, and as a conceptual *requirement* for general science, rather than as something "given" in experience (Buchdahl 1969, p. 476).

The idea that science generally looks for causes of natural phenomena, that could be understood as regular antecedents to these, and that there is a principle (or universal law) of causation—in nature *or* as a conceptual requirement for science—was rejected by Bertrand Russell in his "On the Notion of Cause" (Russell 1912). Russell there formulated several problems for the regularity theory. (For a good recent overview of Russell's paper, see Hitchcock 2007.) He went on to claim that laws of nature as they are actually articulated in modern, mathematical sciences—with astronomical physics as his example—state, *not* the causes of types of phenomena, but functional dependencies between configurations of physical systems at different instances in time.

> In the motions of mutually gravitating bodies, there is nothing that can be called a cause, and nothing that can be called an effect; there is merely a formula. Certain differential equations can be found, which hold at every instance for every particle of the system, and which, given the configuration and velocities at one instant [...] render the configuration at any other earlier or later instant theoretically calculable. (Russell 1912.)

Russell argued that the notion of cause, that he took to be employed in philosophy at that time, is incoherent, and furthermore that mature sciences have grown out of its use, and that it should therefore be purged also from the philosophical lexicon. (More recently, a similar position has been defended by John D. Norton (2003).)

One can dispute Russell's claim that mature science does not employ a cause concept. Patrick Suppes provided contrary examples specifically from physics. He cited seven then-recent article or book titles that mentioned causality (Suppes 1970, p. 5-6). Nancy Cartwright, on the other hand, has argued that causal laws indeed cannot be reduced to regularities, but that they are also indispensable in science, as they constitute recipes for "effective strategies" (Cartwright 1983, p. 22). If so, then causes are needed especially in applied sciences, such as medicine, and in economics and the social sciences, whose results underlie policy decisions.

Some manipulationists have accepted the claim that objective nomic relations between natural events are not causal, but nevertheless wanted to defend the meaningfulness of the concept of cause. (I.e., they are causal anti-realists, of some kind.) They do this by arguing that the notion of cause is rather derived from the experience, and consequent idea, of human influence. As such

it can be retained as a meaningful notion within the sphere of practical concerns and interests, while not necessarily being fundamental to our scientific understanding of the objective physical world.

Not every manipulation theorist rejects causation as part of the objective, physical world. Recent interventionist theories, in particular, make causation more objective, and less tightly associated with human agency. (I.e., they are more causally realist.)

1.4.3. Establishing the asymmetry of causation. Hume's theory distinguished cause from effect by the temporal order of the constantly conjoined event types, and Reichenbach also relied on this time-ordering condition in his probabilistic theory of causation (Reichenbach 1956).

The condition that a cause precedes its effects in time has been seen as problematic by some philosophers. For one, we may attempt to analyze the direction of *time* by reference to the asymmetry between cause and effect. The main reason could be that this causal asymmetry is an objective feature of the world in modern physical theory—Special Relativity specifically—while time orderings are not. (The asymmetry I have in mind is just the one implied by the fact that, according to Special Relativity, if a *signal* can be sent from event `a` to event `b`, then `a` is in the past of `b` relative to every observer.) Along with such a theory, we might hope to explain our *perception* of past, present, and future in terms of the asymmetry between cause and effect. (For an influential suggestion of this kind, see Mellor 1998.)

Some philosophers have been skeptical of the time ordering condition on different grounds. They have thought that claims about time-reverse or simultaneous causation *make sense*, and it can therefore not be an *a priori*, conceptual truth of causation, that a cause always temporally precedes its effects. At most, this would be a contingent truth about our world, belief in which would require some evidence. As David Lewis puts it, we should not reject *a priori* "legitimate physical hypotheses that posit backward or simultaneous causation" (Lewis 1987b, p. 170). If the time-ordering condition is rejected, then a successful analysis of the causal asymmetry must employ other means. A relatively recent, comprehensive treatment of this problem is Hausman 2008.

The strategy of the manipulationists, as regards this particular problem, is now to claim that `A`s cause `B`s only if we can affect the `B`s by manipulating the `A`s, and that this latter relation is inherently asymmetric. This may be historically the most common application of the manipulationist approach to a known difficulty in theories of causation, and might have appeared for the first time in Frank Ramsey's "Law and Causality" (1978, p. 146).

1.4.4. Excluding spurious correlations. In his discussion about causal laws, and their identification with constant conjunctions, Mill introduced a distinction between properly causal regularities and "sequences, as uniform in past experience as any others whatever, which yet we do not regard as cases of causation, but as conjunctions in some sort accidental" (Mill 1882, book III, ch. V). Thomas Reid had provided an example, when he objected to Hume's regularist account of causation, by pointing out that night regularly

precedes day, but that we nevertheless don't think that night causes day. Mill gave the following interpretation of the necessary connection, that he thinks distinguishes causal from accidental regularities.

> This is what writers mean when they say that the notion of cause involves the idea of necessity. If there be any meaning which confessedly belongs to the term necessity, it is unconditionalness. That which is necessary, that which must be, means that which will be, whatever supposition we may make in regard to all other things. (Mill 1882, book III, ch. V.)

A way of understanding this, then, is that night is not the cause of day because their successive occurrences are *conditional* on a third thing, which is the rotation of Earth around its axis.

This suggestion gets a particular treatment in Reichenbach's probabilistic theory of causation (Reichenbach 1956). Reichanbach's *Common Cause Principle* states that any correlation between two event types A and B is explained either by one of them being a cause of the other, or by A and B having a common cause. Excluding accidental regularities—or rather spurious correlations—is then a question precisely of identifying common causes. For doing this, Reichenbach introduced a second assumption, beside the Common Cause Principle, which we call the *screening off condition*. This assumption entails that an event is probabilistically independent of all events that are not among its effects, conditional on its immediate causes. These principles together imply that if A and B are probabilistically dependent, and one is *not* the cause of the other, then there exists an event of type C such that conditional on it, A and B are independent. Thus, if C is the common cause of A and B, then formally, $Pr(\mathtt{B}|\mathtt{A}) > Pr(\mathtt{B})$ but $Pr(\mathtt{B}|\mathtt{A.C}) = Pr(\mathtt{B}|\mathtt{C})$. But note, again, that if A and B are in fact causally related, and C is rather an *intermediate* event, occurring temporally between them, then, in this situation too, A and B will be independent conditional on C. So, again we must refer to the time ordering of the events to find a possible common cause of A and B. We can thus state a probabilistic analysis of causation as

> **PT:** a is a cause of b *iff* $Pr(\mathtt{A}|\mathtt{B}) > Pr(\mathtt{A})$, a temporally precedes b, and there exists no event c such that it temporally precedes a and $Pr(\mathtt{A}|\mathtt{B.C}) = Pr(\mathtt{A}|\mathtt{C})$.

This theory then accommodates cases where a type of cause is not invariably followed by a possible effect of it, as in the case of smoking and lung cancer. Although this account of causation is compatible with fundamentally indeterministic laws, as may be present in particle-scale events according to modern physics, it does not *impose* indeterminism on the causal relation. The probabilistic treatment is compatible with deterministic laws, in which case indeterministic dependencies can be explained rather in terms of varying background conditions. This means in particular that the probabilistic theory treats causation by way of deterministic laws, and the corresponding deductive inferences of effects, as a special limiting case, making the deductive covering-law theory redundant in principle. But note that the explanation of indeterministic dependencies in terms of deterministic laws implies that any event b that is an

indeterministic effect of an event a must have some further causes, in addition to a, that may occur or not occur when a occurs (so that b may occur or fail to occur when a occurs). This assumption of multiple independent causes is also required for the general adequacy of **PT** in identifying causes: in a perfectly deterministic system where c temporally precedes and *alone* causes a, which in turn *alone* causes b, c completely determines whether b occurs or not, so $Pr(\mathtt{B}|\mathtt{A.C}) = Pr(\mathtt{B}|\mathtt{C})$, making b independent of its *direct* cause a conditional on its *remote* cause c. Under such circumstances, the second conjunct in **PT** is not satisfied, and the theory fails to identify a as a cause of b. The way in which **PT** excludes common causes depends, therefore, on that effects are not fully determined by their designated causes, either due to fundamentally indeterministic laws, or the presence of multiple independent causes of any event.

However, it's still not clear that this theory can identify all cases of accidental regularities. In particular, under an empiricist interpretation of probabilities that identifies a probability with an actual frequency, the theory can't identify cases where the common cause is part of the *initial conditions* of the world, since these only occur once, and therefore can't meaningfully be conditioned on. Thus, if it is a consequence of such initial conditions, and not of a causal law, that there is a correlation between a sphere being made of gold, and it being less than one mile in diameter, then this is a case of spurious correlation not excluded by **PT**'s conditions. (This example of a spurious correlation is from Van Fraassen 1989, p. 27.)

The manipulationists' approach to the problem of identifying accidental regularities and spurious correlations is—when this problem is addressed—a correlate of the previous one: if As and Bs regularly co-occur, not because one is a cause of the other but due to having a common cause, then manipulating one would *not* be accompanied by a change in the probability of the other. We can make the additional observation, then, that this condition seems to depend on a counterfactual view of the connection between cause and effect, where the counterfactual involves a manipulation or intervention. The issue of excluding common causes is treated in most technical detail, and in the most overtly counterfactual terms, in recent interventionist theories.

1.4.5. Other problems. Other problems than these are addressed by some manipulation theories. In particular, later theories address problems that have come to the fore in the philosophy of causation only more recently, such as those of causal preemption and overdetermination. I will not be discussing these issues at any length in the monograph, as focusing on the more basic issues will suffice for my purposes.

CHAPTER 2

R. G. Collingwood: Agent relative causation

R. G. Collingwood was a philosopher of history who notably influenced the debate about the nature of historical explanations that followed the publication of Hempel's Deductive-Nomological theory of explanation (see D'Oro and Connelly 2015). In "On the So-Called Idea of Causation" (Collingwood 1938), and especially in his *An Essay on Metaphysics* (Collingwood 1940), he proposed an interpretation of the cause concept in terms of agents' ability to, and experience of, influencing their world. Apart from the fact that several ideas that are prominent in later manipulation theories of causation—especially that of G. H. von Wright—are introduced here, what mainly makes Collingwood's treatment interesting is its radical agent relativism. While this agent relativism could be seen as the whole point of Collingwood's approach to the issue of causation, it is a property that later manipulation theories have, to a successively greater extent, tried to eliminate. This, as we shall see, has not only made the later theories more acceptable to those expecting causation to be a more objective, agent independent, relation in the world, but also introduced new problems.

Collingwood believed "cause" to be an ambiguous term with at least three senses, that he labeled "Type I," "Type II," and "Type III," but which I will call "cause$_1$," "cause$_2$," and "cause$_3$" here, for brevity.

2.1. Cause$_1$: the historical sense

Collingwood calls cause$_1$ the historical sense of the word, and this seems apt for two reasons. Firstly, it is the sense that Collingwood claims is used by historians when explaining historical events "unless they are aping the methods and vocabulary of natural science" (Collingwood 1940, p. 286). As this suggests, the relata of causation$_1$ are particulars. Finding a cause$_1$ is moreover not a matter of identifying a law that the cause and its effect fall under, but rather involves a hermeneutical process of interpretation. Secondly, "cause$_1$" is, according to Collingwood, the *original* sense of "cause," that have imbued all later applications of the concept with suggestions of necessity and compulsion.

Collingwood states that in a claim "`c` caused$_1$ `e`," both "`c`" and "`e`" denote human activities, and

> that which is caused is a free and deliberate act of a conscious and responsible agent, and 'causing' him to do it means affording him a motive for doing it. (Collingwood 1940, p. 290.)

Some agent A causes another agent B to perform some act `e` when they persuade or compel B to perform it, by introducing some states of affairs, or

bringing some states of affairs to B's attention, such that, given B's goals and preferences, B will become convinced that e is what they want to do, and they intentionally do it. e is thus a free and intentional act by B—the force involved in causation$_1$ is purely rational, not physical. According to Collingwood, this is the original—and the only coherent—concept of force or compulsion associated with causation, and in this respect he agrees with the view common at the time, that there are no necessary connections between natural events.

In a further analysis, Collingwood identifies two parts to a cause$_1$: an objective state of affairs, which he calls the *causa quod*, and an intention in the agent (not just a desire), of performing the effect, which he calls the *causa ut*. Importantly, if an agent convinces *themselves* of what are appropriate actions, without the involvement of other persons, then their actions can properly be called *causa sui* (self-caused). Collingwood relates this directly to *responsibility*. *Causa sui* actions entail full personal responsibility for the actions by the agent (Collingwood 1940, p. 294-295).

While currently popular versions of manipulationism live squarely within the paradigm of quantitative science, a hermeneutical and teleological view of actions similar to Collingwood's is central also in von Wright's manipulationist account of causation.

2.2. Cause$_2$: causes as handles

Collingwood associates a different sense of "cause" to the "practical sciences." These he defines as those general scientific pursuits that do not have practical applications as a fortunate side-effect of a search for a purely theoretical understanding of phenomena, but as an essential goal, and he sorts medicine and engineering under this label. The relata of causation$_2$ are natural events (types), such that the cause is "immediately under human control" while the effect "is not immediately under human control but can be indirectly controlled by man because of the relation in which it stands" to the cause (Collingwood 1940, p. 286). We shall return in a moment to what sort of relation this is. As finding such causes is a matter of finding general ways of manipulating events in nature, the relata of causation$_2$ are universals—I call them event types—rather than particulars (Collingwood 1940, p. 308). Collingwood defines causation$_2$ as follows.

> **Causation$_2$:** "A cause is an event or state of things which it is in our power to produce or prevent, and by producing or preventing which we can produce or prevent that whose cause it is said to be" (Collingwood 1940, p. 296-297).

Causation$_2$ is that of Collingwood's senses of "cause" that most closely resembles the main object of later manipulation theories. There are three important differences between Collingwood and later advocates. One is the degree to which Collingwood's theory is agent relative, not only at the species level, but at the level of individuals. The second is the absence of an account of the *logical* role of manipulation in the theory. (That Collingwood does not address the particular problems that were identified especially in connection with the logical treatment of the covering-law theory of causal explanation is unsurprising,

as his theory is published a few years before Hempel's first public account of the D-N model (1942).) The first feature is a consequence of the intimate connection between being a cause$_2$ and being the means by which some agent can accomplish certain individual, practical goals. Thus, it is an immediate result that there is no such thing as a cause$_2$ which could not be manipulated, and this is the third central difference compared to later theories. In fact, Collingwood considers such an idea to be a contradiction in terms (Collingwood 1940, p. 299). And:

> [T]he question whether the effect can be produced or prevented by producing or preventing the cause is not a further question which arises for persons practically interested when the proposition that (for example) malaria is due to mosquito-bites has been established; it is a question which has already been answered in the affirmative by the establishment of that proposition. This affirmative answer is in fact what the proposition means. (Collingwood 1940, p. 299-300.)

Consequently, Collingwood claims that if a clinical scientist purported to have found the cause$_2$ of some ailment, but this causal fact did not imply an effective treatment of that ailment, then the discovery would not be accepted by the medical community, but "denounced as a sham" (Collingwood 1940, p. 300). Independently of whether this is a correct statement about actual clinical research standards, it is easy to understand why later manipulation theorists have wanted to avoid such implications from their theories, as it seemingly excludes the possibility of identifying, say, untreatable genetic causes of some pathology, which surely is a thing.

Moreover, as different individuals differ in what they can do, and also in what they are interested in accomplishing, they will, and *should*, differ in their causal$_2$ judgments. The fact that Collingwood thinks that only those who have some vested interest in producing or preventing some kind of event can form an opinion about its causes$_2$ leads him to reject Hume's explanation of the impression of a necessary causal relation between events. Collingwood believes that passive observers form no causal beliefs at all—Hume "was trying to to explain how something happens which in fact does not happen" (Collingwood 1940, p. 307). The impression of necessity is rather derived, as we have already noted, from the cause$_1$ concept, when this human experience of compulsion in action is projected onto the natural world (Collingwood 1940, p. 309).

The issue of unmanipulable causes is, then, not a problem for Collingwood's theory, as their existence is rejected outright on conceptual grounds. It is rather a problem *with* the theory, for those who are unwilling to accept its radically agent relative implications, or who think they simply make the theory extensionally inadequate. Efforts to amend this situation will commence with Gasking's theory, which we will get to in the next chapter. For Gasking, as for the manipulation theorists who follow him, the existence of *prima facie* unmanipulable causes is an issue that requires some theoretical response.

Besides his radically agent relative conception of causes$_2$, Collingwood relates these causes to Mill's theory of causation in a way that at first glance

makes them seem considerably less controversial than they may otherwise have appeared to be. He notes that causes$_2$ are always dependent for their effects on further conditions—what we usually call the relevant background conditions. Mill had proposed that the "real cause" is that which is invariably and unconditionally followed by its effect, that is to say the totality of all and only the conditions required for the effect—what we may call the total cause. Now the condition that is singled out in a particular causal claim (the "designated cause") is rarely if ever this total cause, but just some part of it. Collingwood proposes that the condition that is designated as the cause in a causal claim is selected based on what "I am able to produce or prevent at will" (Collingwood 1940, p. 302). He now states his principle of the relativity of causes:

> [F]or any given person the cause in sense II [cause$_2$] of a given thing is that one of its conditions which he is able to produce or prevent. (Collingwood 1940, p. 304.)

It is a further corollary to this principle that if there are no such manipulable conditions of the event in question, to some person, then the event has no causes$_2$ for that person (Collingwood 1940, p. 306). So, when Collingwood claims that causes$_2$ depend for their existence on "human volition" (Collingwood 1940, p. 313), it would seem as though it is their existence *as causes* that is so dependent, not their existence as nomic conditions for the effect.

In other words, while causes$_1$ indeed must be understood as irreducibly anthropomorphic things, existing in the realm of reasons, not laws of nature, and it is these causes$_1$ that bring the impression of modal force to the idea of causation in general, causes$_2$ on the other hand seem by Collingwood's description to after all be understandable to an interesting extent in terms of a covering-law theory, together with a pragmatic rule for the selection of that nomic condition for the effect which is properly called "the cause." Collingwood, on this interpretation, agrees that Mill's nomic conditions are *conditions* for the effect, but not that they individually are *causes* of it. Mill, of course, would not object to this. At this point in Collingwood's argument, the difference between Mill and him boils down to whether it is the full set of necessary conditions for the effect, or the pragmatically salient such condition, that ought to be called "the cause." But we shall see in the next section that Collingwood also regards Mill's real causes as impossible, because he thinks that their description is self-contradictory.

It is widely acknowledged that the grounds for selecting one factor over the others in a singular causal claim are pragmatic, and relate to the explanatory context. For example,

> most causal realists are prepared to allow that pragmatic principles of 'invidious selection', as Lewis calls them, govern the way in which we select as 'the cause' a salient part of the vast network of events leading up to an event. (Menzies 2007, p. 2.)

(See also, e.g. Lewis 1987a, p. 215-216). Generally speaking, the choice is determined by such things as what is perceived as abnormal or unexpected,

as manipulable, as relevant to assignment of blame or credit, or as salient for other reasons in the context of some question that prompted the causal claim.

Relative to later theories that identify *a* cause with some nomically necessary condition for the effect, leaving the designation of any particular such condition as *the* cause to contextual salience, the difference to Collingwood's treatment seems then to be just that his pragmatic considerations are restricted to practical ones, and that he doesn't acknowledge that the multitude of conditions of some effect are all causes of it. We might then say that, from one point of view, the sort of relativism expounded by Collingwood is something that pretty much all later theories of causation have embraced, in the form of a relativity to differing interests in a broader sense than Collingwood's.

More recent manipulation theories employ manipulations for the purpose of excluding cases of spurious correlations, but Collingwood does not put things this way. He appears to accept that Mill's real causes don't include cases of accidental regularity. Another point at which later manipulation theories rely fundamentally on manipulations is in establishing the *asymmetry* of the causal relation. But Collingwood does not employ manipulability expressly for this theoretical purpose either—although the fact that he does comment on the problem of the "priority" of cause to effect in connection with his final sense, causes$_3$, may suggest that he takes manipulability to do that job in the two previous senses.

Despite the broadly unpopular aspect of radical agent relativism in Collingwood's theory of causation$_2$ (see e.g. Woodward 2014a, Hausman 1997), it has several parts that resemble and inform later manipulation theories. Collingwood clearly articulates the view that causal$_2$ knowledge is knowledge about how to go about accomplishing practical things, an idea that has been adopted in such terms as "recipes" (Gasking 1955) or "effective strategies" (Cartwright 1979).

> If sciences are constructed consisting of causal propositions in sense II [cause$_2$] of the word 'cause,' they will of course be in essence codifications of the various ways in which the people who construct them can bend nature to their purposes, and of the means by which in each case this can be done. (Collingwood 1940, p. 307.)

Collingwood moreover claims to have analyzed the cause$_2$-effect relation in terms of a means-end relation (Collingwood 1940, p. 308)—much the same claim that Gasking will later make.

Collingwood also gives a description of that action that is the manipulation, and its relation to what is directly manipulated, that strongly resembles descriptions of actions and their results in later manipulation theories.

> Turning a switch to one or other position by finger-pressure is an instance of producing a certain state of things (the ON or OFF position of the switch) *immediately*, for it is nothing but a certain complex of bodily movements all immediately produced. These movements are not our *means* of turning

> the switch, they *are* the turning of the switch [my emph.]. (Collingwood 1940, p. 297.)

Von Wright in particular elaborates on this idea in his theory of action. That the manipulation does not *cause* its immediate result—as it would on Collingwood's account if the bodily movement was the *means* by which the switch is turned—is important. If manipulations are instances of causation, then the manipulation theory can at least not be an analysis or definition of causation in other terms, on pain of circularity.

2.3. Cause$_3$: objective causation

The final sense of "cause" that Collingwood proposes relates natural events independently of practical concerns, in a purely theoretical account of phenomena. This is Mill's "real cause," a concept presumed to occur mainly in the natural sciences in their theoretical forms. Collingwood ascribes the following properties to the relation denoted by "causes$_3$":

> (a) if the cause happens or exists the effect also must happen or exist, even if no further conditions are fulfilled, (b) the effect cannot happen or exist unless the cause happens or exists, (c) in some sense which remains to be defined, the cause is prior to the effect; for without such priority there would be no telling which is which. (Collingwood 1940, p. 285-286.)

Collingwood goes on to suggest that the priority may be temporal or something else entirely, without offering specific suggestions. He contrasts "cause$_3$" to "cause$_2$" mainly in terms of causes$_3$ being unconditionally sufficient as well as necessary for their effects, while causes$_2$ are causes only given additional *sine qua non* conditions. He presents two problems for causation$_3$: (i) it is a "one-one" relation, which he argues is incompatible with the popular idea that causes temporally precede their effects, and with causes and effects being spatiotemporally distinct in general; and (ii) it cannot make sense of its inherent necessity.

The argument for (i) goes as follows. If $\mathtt{a}_t$ is the cause$_3$ of $\mathtt{b}_{t'}$, t and t' being their times of occurrence and $t'-t=\tau$, then if the time interval τ between $\mathtt{a}$ and $\mathtt{b}$ is not 0, certain things must hold in that time interval, and these are *sine qua non* conditions for $\mathtt{a}$'s causing $\mathtt{b}$. Thus, if there are no *sine qua non* conditions, it must be that $\tau = 0$, and cause and effect are simultaneous. The argument for that $\mathtt{a}$ and $\mathtt{b}$ must occupy the same spacetime region is the same, *mutatis mutandis* (Collingwood 1940, p. 314-315). This is clearly a misunderstanding on Collingwood's part. It is true that certain things must happen or be the case in the time interval between $\mathtt{a}$ and $\mathtt{b}$, under the ordinary assumption that the propagation of effects doesn't "skip over" some parts of spacetime. But that these things do occur can be ensured in a deterministic system by conditions holding at t, the time of the cause. To say that the cause is unconditional is to say that it itself ensures those intermediate things, that is, it includes those conditions holding at t. That Mill's unconditionally sufficient "real cause" must include sufficient conditions for the whole causal chain of events between $\mathtt{a}$ and

b is precisely what makes it something different from our ordinary concept of cause, just as Mill emphasizes.

As to (*ii*), Collingwood presents three interpretations of the necessity in causation$_3$: (*a*) it may be taken to be logical, as was done by the rationalists (he includes Russell's interpretation of "necessity" as the necessary truth of a propositional function here); (*b*) it may be identified with observed regular co-occurrence, as by the Humean empiricists; or (*c*) it may be associated rather with an anthropomorphic sense of "necessary," as in a compulsion to act in some way.

Collingwood argues that both (*a*) and (*b*) are descriptively false, as to the intended meanings of actual language users and in particular scientists. By "a causally necessitates b" they do not mean "logically implies"—whatever it is, knowing about it requires empirical investigation. They also do not mean mere observed regular co-occurrence—this is a confusion of the meaning of the claim with its evidence (Collingwood 1940, p. 316-321).

It is (*c*) that, according to Collingwood, gives the right sense of the necessity involved in causation, as regards actual language use. His argument in support of this is a quite elaborate theory of how an anthropomorphic notion becomes projected onto the objective physical world, by way of theology. To skip over it is in a way an injustice, because the investigation into the historical uses of the concept *is* the metaphysical investigation, according to Collingwood's overarching theory of metaphysics and philosophical method. Since my interest is mainly in the parts of Collingwood's theory of causation that are somewhat continuous with the other theories in our story, I will nevertheless abbreviate his view here. Historically, people have tended to project intentions and agency onto natural events, as exemplified by animistic religions, and teleological explanations of such events in general. However, this anthropomorphic sense of necessity in nomic relations between natural events, that do not essentially involve the interests and capacities of agents, while psychologically explainable, is a myth. It is not there, and modern physics, from Newton and onward, recognizes this (Collingwood 1940, p. 322-327). On this latter point, Collingwood appears to rely on Russell's argument in Russell 1912.

Thus, the necessity inherent in causation is anthropomorphic and this makes sense in the sphere of human reasons and actions, as in causation$_1$, and also in the context of the practical sciences and causation$_2$. But Collingwood considers "causation$_3$" to be both inconsistent and its traditional connotations of necessity archaic and false. He agrees, then, with Russell in that the theoretical sciences neither do nor should employ the notion of cause. But, *pace* Russell, Collingwood does not think that, when social scientists or engineers use this notion, it is a sign of a "backwards" discipline, but rather of a practical one.

2.4. Conclusions: manipulation in Collingwood's theory

Collingwood's causation$_1$ belongs to the hermeneutical tradition of interpreting reasons and motives for actions. It thus has nothing to do with an

objective relation between natural events. This he also thinks is the true domain of causal necessity and force—as he takes these notions to be grounded in particular human experiences of being rationally compelled to act in a certain way. This sense of causal force can be extended to causes$_2$, which are some means by which an agent can indirectly control events in nature, in accordance with the agent's purposes. These causes$_2$ are agent relative at the level of individuals, but only in the sense that they denote those objective nomic conditions that are of practical interest to that individual. That the causal factor that is designated as *the* cause in a particular causal claim is determined on pragmatic grounds is widely acknowledged across different types of theories of causation, in a way not wholly unlike that of Collingwood. Finally, Collingwood argues that "`a` causes$_3$ `b`" is incoherent under the Millian understanding of "cause," but this argument appears to come out of a misunderstanding of the nomic analysis, at least under an assumption that causally related events are governed by deterministic laws. And he also believes that causation$_3$ *falsely* assigns an anthropomorphic sense of necessity to an objective relation between natural events. As to this last point, it is clear that rather than advocating a problematically anthropomorphic sense of causation, Collingwood would take himself to be in the business of *eliminating* anthropomorphism from the domain of natural laws.

Collingwood does not explicitly engage with any of the logical problems identified in later treatments of causation, such as identifying the direction of causation or excluding spurious correlations, and accordingly does not enlist manipulations to resolve them.

CHAPTER 3

Douglas Gasking: Manipulation and Causal Asymmetry

Douglas Gasking outlines a manipulation theory of causation in a short paper in 1955, titled "Causation and Recipes" (Gasking 1955). Neither Gasking nor any of the later manipulation theorists we will cover presents or employs anything like Collingwood's cause_1 , that denotes causation only between human actions. The relevant comparison is to Collingwood's cause_2 . There are however two central differences between Gasking's causes and Collingwood's cause_2 . Firstly, Gasking aims his theory specifically at what he sees as the inadequacy of the covering-law theory's account of the asymmetry between cause and effect—the fact that if `a` is the cause of `b`, then `b` is not the cause of `a`. In the covering-law theory, this was established by requiring that causes precede their effects in time, and Gasking presents what he takes to be counterexamples both to the necessity and the sufficiency of this condition. Secondly, Gasking's causes and effects extend to natural events that may not be manipulable, by human beings or other agents. That is, his ambition is to give a manipulationist account of a general causal relation between natural events—and this ambition is shared by all later manipulation theorists we shall encounter. All the same, Gasking's view is that causation is grounded in, or derived from, ways in which we as agents can indirectly control parts of our environment—what he compares to "recipes."

Gasking means that the notion of cause that is related to recipes in this way "is the fundamental or primitive one," and not employed by scientists in their careful formulations, except by those who are involved in decidedly practical investigations, such as "engineering, agriculture or medicine." Advanced science expresses their findings rather in terms of pure functional laws, that license inferences (Gasking 1955, p. 486-487). This, then, sounds similar to Collingwood's restriction of the proper application of the cause_2 concept, but without the relativity to individual agent goals. But Gasking importantly claims that we can extend proper causal claims also to events that could not be manipulated.

3.1. Suggested counterexamples to the time ordering condition

As a reminder, the most popular theory at the time required that effects can be derived from their causes together with laws. Since, in a deterministic system, it is possible also to derive the cause from the effect, there was the

further condition that causes precede their effects in time. Gasking argues that these conditions are inadequate for identifying causes.

3.1.1. Against the necessity of time ordering. As the temperature of a piece of iron increases, it will eventually start to glow. This is a law of physics. It will certainly glow at a 1,000º C. If we now asked what causes the iron to glow, we should answer, says Gasking, that it glows because its temperature is at least 1,000º C. But the glowing is not an event subsequent to the temperature increase, they are simultaneous. Thus, the time ordering condition in the covering-law theory is not satisfied.

In another example, we can increase the resistance in a circuit, and thereby decrease the current flowing through it. The change in current can reasonably be said to be caused by the increased resistance. But again, the decrease in current does not occur after the increase in resistance, but simultaneous to it, so once more the time ordering condition is not satisfied.

3.1.2. Against the sufficiency of time ordering. Using the law of gravitation we can calculate the velocity of a free-falling object o at a time t_2 from its velocity at an earlier time t_1 in its fall. Hence, the covering-law conditions are satisfied. "But it would be a most unnatural and 'strained' use of the word 'cause' to say that" the object's velocity at t_2 was caused by its velocity at t_1 (Gasking 1955, p. 480). This he takes to show that the time-ordering condition is insufficient.

Moreover, Gasking points out that if the problems in the previous section led us to modify the time ordering condition, and say that a cause either precedes or is simultaneous to its effect, then the glowing of a piece of iron would qualify as the cause of its temperature, and the current flowing through a circuit as the cause of its resistance, but both of these consequences sound intuitively unacceptable to him, and the condition could therefore again be seen to be insufficient.

3.2. Gasking's theory

Having concluded that the covering-law conditions of inferrability and time ordering are inadequate for telling causes from their effects, Gasking proposes a theory that connects causation to manipulation and agency by distinguishing cause and effect based rather on what "manipulative techniques" are available to us. He suggests that we say that heat causes glowing in a piece of iron—and not the other way around—for the following reasons. We possess a general manipulative technique for making things hot, namely putting the thing on a fire. We have *no* general manipulative technique for making things glow. In particular, the technique for making things hot does not make *everything* it is applied to glow, with water as the example to support this claim. But the technique for making things hot *does* make pieces of iron glow, and there is no *other* technique for making a piece of iron glow. Thus, heat causes pieces of iron to glow, and not *vice versa*. This is due then to the asymmetry in the available manipulative techniques: we can make things glow only by using the general technique for making them hot, but not all things treated in that way start to

glow. Gasking also describes an alternative world, in which agents do have a general manipulative technique for making things glow, and this technique also makes iron, but not all things, hot, and he claims that if ours had been such a world, then we would have considered glowing to be the cause of heat in pieces of iron.

If, on the other hand, our only available manipulative technique invariably brought about *both* heat and glowing, Gasking claims that we would not then say that heat was the cause of glowing, nor glowing the cause of heat. That is, if a manipulative technique invariably produces two things, one will not be taken to be the cause of the other. He illustrates by a clever example: as an object is always simultaneously round to the touch and visibly round, so that we cannot make an object exhibit one and not the other of these perceptually distinct properties, we would not say that the object's visible roundness caused its tangible roundness, or *vice versa* (Gasking 1955, p. 485-486). In addition, he takes it that it is the constant co-occurrence of these types of perceptions that explains why we do not have two different words for them. This he takes to also explain the example previously given, of how the velocity of a free-falling object at one time is not said to cause its velocity at a slightly later time. There is no manipulative technique that will bring about the former state and not the latter. That is, dropping something from a sufficient height will make it accelerate in the same way until it hits the ground, whatever it is that is dropped. Thus, there is in this case no distinction among our manipulative techniques on which to ground a distinction between cause and effect.

We can summarize Gasking's conditions and derive some important consequences from them. First, we want to make some simple amendments. Gasking mentions as an independent condition that there is no general manipulative technique for producing the effect, but this follows from the other conditions. In addition, the condition that there is a *unique* manipulative technique for bringing the effect about in the restricted class of things (glowing in iron, in Gasking's example) seems unnecessarily strong, and invites trivial counterexamples. (We can make iron glow by passing a sufficiently strong current through it, or by compressing it with sufficient force. Both of these are also ways of making the iron hot.) It ought to be sufficient for Gasking's purposes that *every* manipulative technique that produces the effect in the restricted class also produces the cause, and that it satisfies the other conditions.

For precision and clarity, we can now put the amended conditions in symbolic form. Let the predicates have the following interpretations: Mx *iff* x is a "manipulative technique," or "method" for short; Axy *iff* x is applied to y; and C, E, I are distinct natural empirical properties. $K(C, E, I)$ should be read as a general causal claim "C is a cause of E in Is."

	$K(C, E, I)$ *iff*	*Translation*
1.	$\exists x($	There exists
	$Mx \wedge$	a method, such that
	$\forall y(Iy \wedge Axy \rightarrow Ey)$	every I it is applied to is E,
	$) \wedge$	and
2.	$\forall x(Mx \wedge \forall y(Iy \wedge Axy \rightarrow Ey) \rightarrow$	every such method is such that
	$\forall y(Axy \rightarrow Cy)$	everything it is applied to is C,
	$) \wedge$	and
3.	$\neg\exists x(Mx \wedge \forall y(Axy \rightarrow Ey))$	no method is such that everything it is applied to is E.

For the theory to ensure the wanted asymmetry between the cause C and effect E in Is, it must imply that if C is the cause of E in Is, then E is not the cause of C in Is. Although a logical argument to this effect is not explicitly provided by Gasking, we can show here that it is a consequence of his conditions, based on our formalization of them. If $K(C, E, I)$, then by condition 3, *(i)* no method exists such that everything it is applied to is E. But if $K(E, C, I)$ then by condition 1 there exists a method such that every I it is applied to is C, and by condition 2 this method is such that everything it is applied to is E. So, *(ii)* there exists a method such that everything it is applied to is E, in contradiction with *(i)*. Thus, $K(C, E, I) \Rightarrow \neg K(E, C, I)$, as desired.

In connection with this, we should note that, on this theory, the relation between the method x and its application to some y, that ensures that y is C when $K(C, E, I)$ holds, whatever it is, is *not* causal. Intuitively, the very fact that *all* things to which x is applied are C contradicts the conditions for C being an effect of x—specifically the condition that there exists something in which the cause is present but that is not C. Or, in other words, Gasking takes causes to be generally insufficient for their effects, but the relevant manipulative techniques to be generally sufficient for their immediate result. It is thus an implication of Gasking's theory, just as it was in Collingwood's, that manipulations are *not* themselves instances of causation.

Let's now turn to some problems with Gasking's analysis. Unfortunately, it is unlikely that even Gasking's own example, of heat causing glowing in pieces of iron, satisfies the conditions on K. It is certainly true that not everything that is put on a fire starts to glow. If water is put on a fire, often the fire will go out. If water in an open, heat-resistant container is put on a fire, it will evaporate long before it reaches glowing temperatures. But if water in a closed container, strong enough to hold the heated water in, is put on a sufficiently hot fire, then the water will eventually glow (whether we are able to see it or not). All objects, independently of what they are made of, emit electromagnetic radiation when they are hot, and emit electromagnetic radiation within the visible spectrum when they are sufficiently hot. That is to say, either we do not have a general method by which we can make anything reach a certain temperature—for example because some items will be destroyed

before that happens—meaning that condition 2 is violated, or, if some method makes everything hot, then it also makes everything glow, violating condition 3.

Because of this very tight connection between temperature and the emission of electromagnetic radiation, understanding how heat could cause glowing is hard for any theory of causation. (For example: the simultaneity of the supposed cause and effect are a problem, as we saw, for theories depending on time ordering for the causal asymmetry. An analysis in terms of David Lewis's counterfactual theory, on the other hand, would imply that, for some actually glowing piece of iron, in a world that is among the most similar to the actual one but in which that iron does not glow, it would still be hot (Lewis 1987b, p. 38 and Lewis 1987c). But how similar would that world really be to ours?) The problem might be that the proposed relata approach denoting the same thing, in two different ways. That is, the heat and the glowing are not *distinct* enough to be causally related. But Gasking is clearly right in that "The iron glows because it is hot" looks like an explanation, while "The iron is hot because it glows" does not. I won't suggest an account of this appearance here, but I will point out that, while there is no contrast among real objects between hot ones that glow and hot ones that don't glow, there is a contrast between glowing objects that are hot and glowing objects that aren't—such as light emitting diodes (LEDs), fluorophores, and plutonium. So a piece of iron, or a wire filament in a light bulb, may glow because it is hot, while this is not the right explanation for why an LED glows. This cause, heat—if it is a cause—is in other words, and in exact contradiction to what Gasking's theory expects, sufficient but not necessary for the glowing effect.

We can show that, in general, the insufficiency of the cause for its effect cannot deliver the asymmetry we want. This is because the insufficiency, too, is in reality symmetric. Just as the designated cause is often sufficient for the effect only given further conditions, the effect usually determines its cause only given further conditions. In terms of Gasking's example—and ignoring the particular difficulties with it that we have already discussed—only if a glowing object is something other than a light-emitting diode, a fluorophore, plutonium, etc., is that glowing sure to be caused by the object being hot. In other words, causes are ordinarily neither unconditionally sufficient, nor—as Gasking claims—unconditionally necessary, for their effects. But it is precisely the fact that causes are taken to be generally necessary and insufficient for their effects that ensures the asymmetry of the relation in Gasking's theory .

The fact that, under Gasking's theory, causes have a general manipulative technique associated with them, such that it realizes the cause in anything it is applied to, but precisely the *opposite* is true of effects, also has the consequence—fatal for the theory under perfectly ordinary expectations—that nothing that is an effect of something can be a cause of anything. There are thus no causal chains of events.

Perhaps surprisingly, then, Gasking's solution to the problem of the causal asymmetry has nothing to do, in the end, with any properties of manipulations.

It is a consequence solely of the logical conditions imposed on the causal relation. However, the theory is still a manipulation theory, because—in the first step—it implies that if something is a cause, then there is a general manipulative technique for producing it. Gasking now extends his theory to particular events that we cannot manipulate, such as the heat radiation of the sun onto Earth.

3.3. Extending the relation to unmanipulable causes

Collingwood rejected the idea of causes that are not manipulable as a contradiction in terms, and claimed that the concept "cause$_3$," taken to denote purely theoretical causes, is incoherent. Gasking, while grounding the meaning of causal claims in "recipes" by which we can indirectly produce or prevent some event or state, acknowledges nevertheless that, sometimes, what we want to make is rather "a theoretical point," about something that we in fact could not manipulate (Gasking 1955, p. 483).

> For example, one may say that the rise in mean sea-level at a certain geological epoch was due to the melting of the Polar ice-cap. But when one can properly say this sort of thing, it is always the case that people can produce events of the first sort as a means of producing events of the second sort. [...] We could come rather closer to the meaning of "A causes B" if we said: "Events of the B sort can be produced by means of producing events of the A sort." (Gasking 1955, p. 483.)

(This example—in light of current theories about global warming—goes to show that it can be hard to say what humans can or can't affect by their actions.) Gasking does not elaborate on what the sorts are here, but we might take them to generally be what I call "event types" in the definition of causal manipulationism, and elsewhere denote by "`A`," "`B`," .... The melting of the polar ice caps would be an instance of water changing its state from solid to liquid, and we can certainly produce effects of this type on a smaller scale. The problem is not to find suitable types for the causally related particulars, such that these types have manipulable instances, but to somehow restrict the types, since without principles anything can be classified together with anything else. As it seems an empirical fact, and not a formal or logical one, that heat from the sun changes the state of water on Earth in the same way ice in a pot can be changed to water by applying heat from a stove, we can probably do no better than appeal to a classification into "natural" types of events.

The important point is that Gasking, like later manipulationists, wants to include theoretical causal claims in the manipulation account of causation. This makes the theory more easily digested by someone who accepts that causal claims that are not of immediate utility can still be meaningful and true, but it introduces this new complication as to unmanipulable causes. A strategy broadly similar to Gasking's is employed later by von Wright and also in the paper by Peter Menzies and Huw Price, although the idea is there described somewhat differently and in more detail.

3.4. The charge of circularity

In a 1973 paper aimed at von Wright's then recent proposal, Alexander Rosenberg expresses several objections to Gasking's theory. He considers Gasking's account to be "hopelessly unilluminating" for several reasons (Rosenberg 1973, p. 378). I will focus on Rosenberg's objections that have to do with *circularity*.

Rosenberg claims that circularity infects Gasking's account in three places. Firstly, when Gasking says that we have a manipulative technique for making iron glow—which is the general manipulative technique for making things hot—Rosenberg takes it that "making" *means* "causing."

> But if this is to be Gasking's meaning, then his remarks in no way elucidate the relation of heating and glowing; for he is making the unilluminating statement that if the manipulative technique for heating causes glowing then heating causes glowing. Who would deny this? (Rosenberg 1973, p. 378-379.)

This statement is not analytically true, as Rosenberg seems to suggest here. In fact, the inference is not even valid given ordinary expectations on the causal relation, because the technique that heats things could also make iron glow, but not *by way* of the heating, but by way of some other thing done to the iron, as part of the manipulation for heating. This is a situation where the manipulative technique is itself a confounder, and in which the conclusion that heat causes glowing in iron would be in error. The possibility of this sort of confounder is explicitly excluded only in more recent, intervention theories such as James Woodward's (2003). This, then, points to a kind of situation in which the antecedent in Rosenberg's quoted reformulation of Gasking's proposal is true and the consequent false. This is not to deny that "making" is a causal term, and the definition therefore circular in this respect. My point is just that Gasking's theory does not essentially depend on such a circular formulation. A reformulation can eliminate the circularity, as in "Cs cause Es in Is *only if* there is a method for making things C, and if this method is applied to Is, then these will be E." (That is, the condition that E is realized is stated in terms of a mere co-occurrence.)

Secondly, Rosenberg says that "the relation between the actions which constitute the technique and the iron's becoming hot is causal as well" (Rosenberg 1973, p. 379). This would indeed create a circularity in the analysis, but as I have wanted to emphasize above, whether intentional or not, Gasking's theory logically excludes the possibility that manipulations cause their immediate results.

That the early manipulation theories are circular, due to "manipulation," "producing," "bringing about," etc. themselves being causal terms has become a standard objection to these theories (e.g.: Hausman 1997, p. S17; Paul and Hall 2013, p. 38; Psillos 2014, p. 103; Woodward 2014a, p. 1715). While it is a natural and understandable view that these indeed *are* causal terms, and that a manipulation is a physical event that, if it results in, produces, or brings about something, causes that thing, the earliest manipulation theories nevertheless imply that this is not the case. It seems to me, therefore, that

the standard circularity objection to these theories at the very least gets its emphasis wrong.

Finally, Rosenberg claims that when Gasking extends the causal relation to events that are not manipulable, by way of these events belonging to a class with some manipulable members, it's not possible to define this class without referring to causal relations or properties.

> Consider Gasking's own case: what are the relevant similarities between the events at the polar ice cap and at the bucket of water, which makes them the same 'sort'? I suggest they are causal. (Rosenberg 1973, p. 382.)

Perhaps this is correct and unavoidable for some classes that would have to be defined. But it doesn't seem right to me in this particular case. Both the situation involving the polar ice caps and that of the bucket of water are instances of water changing from its solid to to its liquid state. Although Gasking uses the term "melt" which certainly is a causal term (and he recognizes this), we don't seem to *need* to refer to causal properties to define this class. But it doesn't follow that we can always avoid referring to such properties in the definition of the event types. Perhaps some types of events *are* classified on just those causal properties that are in question. That the relevant similarity between the manipulable and the unmanipulable cases must appeal to causal similarities can be given a more sophisticated argument, that is provided by James Woodward in his critique of Peter Menzies and Huw Price's theory, here presented in section 5.5 (p. 79).

In a response to Rosenberg's criticisms of Gasking's theory, Yehudah Freundlich made a suggestion that has reappeared in more recent times. Accepting that Gasking's account of causation indeed is circular, he claimed nevertheless that "[m]ere circularity in an analysis is not in itself a sufficient reason for deploring that analysis" (Freundlich 1977, p. 475). This is echoed later in James Woodward's defense of his own theory. Woodward connects theoretical circularity with reduction in a familiar way, assuming that a theory that employs circular definitions cannot be giving a reductive account of causation. "[I]t is crucial to my argument that an account of causation and explanation can be worthwhile and illuminating without being reductive" (Woodward 2003, p. 21).

Freundlich also states that

> what is being offered is not a definition of causation in terms of more "primitive" elements, but an analysis of causation in terms of the general means of checking, or testing, for the existence of the causal relation between states. The central problem which the analysis proposes to solve is: wherein lies the asymmetry of a causal relation between A and B which is over and above the mere symmetric lawful relation between them? Causation is thus elucidated by supplying the constitutive conditions for a causal relation to exist (namely, that the cause can in principle be used to bring about the effect, but not vice versa). (Freundlich 1977, p. 476.)

This connects directly a *constitutive condition* for the presence of causation with what is clearly a *sufficient epistemic condition* for detecting it. But the relation between the two is anything but obvious. Rather, broadly speaking, from that A-type events are causes of B-type events and that A-type events are manipulable, it follows that we could discover this general causal relation (when the conditions in other respects are favorable). But this does not make manipulability constitutive of causation—for this claim some further argument appears to be required.

3.5. Conclusions: manipulation in Gasking's theory

Let's revisit some aspects of Gasking's theory. First, and relating to what sort of restrictions the theory imposes on causal claims, if we look back at Collingwood's theory, saying there that As cause$_2$ Bs only if As are a "handle" by which we can bring about or prevent Bs—or rather, the condition is even stronger: by which we can accomplish some *existing practical goal* that involves bringing about or preventing Bs—severely restricts what are valid causal claims, and has radically revisionary consequences, since all purely theoretically causal claims are rejected. This shows that Collingwood's theory makes a very substantial difference to what we correctly call a cause, compared to its competitors. But, due to his extension of the cause concept to unmanipulable events, this particular difference does not seem to be present in Gasking's theory, as it aims to acknowledge also the theoretical causal claims we usual take to be credible. We may then ask: what work do manipulations do in the theory? What is explained, or what problem is solved, by introducing manipulative techniques? Gasking does not engage with the problem of causal necessity. In particular, he does not say that the impression of a necessary relation, or causal force, between events has its origin in the human experience of rationally compelling reasons to act in a certain way. A hermeneutical perspective doesn't explicitly enter into Gasking's picture. And I have argued above that while Gasking succeeds in establishing an asymmetry between cause and effects as a consequence of the conditions he impose in his theory, this asymmetry is not ultimately due to any property of manipulations, and the desired result anyway comes at too high a price. In particular, the way the conditions exclude the possibility of causal chains, and postulate that causes are generally necessary for their effects, is deeply problematic.

Gasking does hint toward a possible feature of manipulations that might serve explanatory purposes. He takes "A causes B" to mean that events of the B sort can be produced by means of producing events of the A sort, and this "fits in with the principle that an event A at a time t_2 cannot be the cause of an event B at an earlier time, t_1. It is a logical truth that one cannot alter the past" (Gasking 1955, p. 483). But even if this is a logical truth, it seems to be implied just by the past being fixed, and thus doesn't depend on any features that are particular to manipulations.

But maybe there is another way altogether of understanding Gasking's proposal. He tends to speak descriptively about what causal claims we actually make, rather than in terms of what causes *are*. Accepting that the theory states

something true about the conditions under which we make such claims, an *epistemic* explanation of this fact suggests itself, according to which the theory should be understood as explaining something about causal beliefs in agents, rather than stating what causation is, or what the content of the concept of cause is. After all, it is plausible that our belief that the heat radiation from the sun causes the presence of liquid water on Earth is based at least partly in practical experiences of making water melt. But on this reading, Gasking's theory is not a special kind of manipulation theory of causation, as has generally been assumed, but a theory about the formation of causal beliefs.

I do not imagine that Gasking, if he knew of the problematic consequences of his conditions that I have detailed here, accepted them, or that he would have if he didn't. When von Wright says that the earlier theory of causation most similar to his own is Gasking's, I can only understand that as a reference to the way Gasking generally connects causes conceptually to manipulations, in a way that is intended to be largely non-revisionary relative to the well founded theoretical causal claims that we actually make. In this respect, Gasking's theory is an important precursor of modern manipulation theories of causation.

I think that the main lesson we might learn from examining Gasking's proposal is that for a manipulation theory of causation to explain something previously unexplained, or to solve some outstanding theoretical problem, it must say something about what manipulations *are*—imbuing them with some specific properties that can do this theoretical work.

CHAPTER 4

G. H. von Wright: Action and Causal Possibility

4.1. Introduction

Von Wright lays out his theory of causation especially in *Explanation and Understanding* (von Wright 1971, ch. II), "On the Logic and Epistemology of the Causal Relation" (von Wright 1973), and *Causality and Determinism* (von Wright 1974). He there recognizes that there may be numerous senses of "cause," and intends his proposal to be applicable just to one of them, namely that one which is importantly involved "particularly in the experimental and natural sciences" (von Wright 1974, p. 1). By way of contrast, von Wright does not mean to say anything about a cause concept such as Collingwood's cause$_1$, where the effects would be human actions. What, in turn, the "determinants" of actions are, von Wright elaborates on in his theory of action and agency.

Recent references to von Wright's theory (e.g. Woodward 2016) have tended to sort it with the theories of Collingwood and Gasking, and focus on the definition of causation that von Wright presents in *Explanation and Understanding*:

> 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.)

Like Gasking, von Wright goes on to extend the causal relation to events that could not be manipulated. His way of doing this is different from Gasking's—von Wright claims that all causes are at least *composed* of events that are manipulable. More importantly, however, this definition by itself does not answer important questions about what sort of work manipulations do in von Wright's theory, and how they do it. The purpose of this chapter is to try to extract some answers to these questions.

There are two parts to von Wright's theory of causation, that are interspersed throughout the texts, but that can nevertheless be untangled in a way that I find helpful for understanding the proposal. I'll call these parts the *logical* and the *conceptual* analysis, respectively. We shall see that von Wright's analysis of causation in terms of action and agency happens mainly in the conceptual part of his theory.

As to the logical part of the theory, the analysis of the logical properties of the causal relation proceeds in terms of a formal, modal language as well as a graphical representation of "possible histories"—or perhaps more accurately histories of possibilities—in the form of a topological tree. Within these two

ways of representing causal systems, von Wright identifies "conditionship relations" between states of affairs in the system, and conditions on a causal relation in terms of these conditionship relations. Von Wright moreover calls these conditionship relations "nomic" and also "causal." This analysis is then superficially quite reminiscent of the traditional analyses of causation in terms of nomic dependencies, that were inspired by Mill, even though von Wright's graphical representation of causal systems and his formal apparatus are novel. But in at least two ways, it is clear that von Wright does not analyze the causal relation in terms of nomic dependencies. Firstly, his theory is supposed to rely somehow on manipulations. Secondly, he sometimes calls the dependencies themselves (i.e. the conditionship relations) that determine what causal relations there are "causal." The latter observation was expressed by Dag Prawitz in Prawitz 1989 (p. 431): "[I]t is clear that [the causal languages constructed by von Wright] take the causal concept for granted. They do not constitute an attempt to understand the concept of cause in terms that do not already presuppose this concept." In his response to Prawitz, von Wright does not address this point explicitly. However, he takes the notion of law-governed possibility to be *conceptually* dependent on a more primitive notion of possible *action*. "My way has been to make the notion of nomicity rest on the notion of causal counterfactual conditionals, and then try to show that this notion of counterfactuality has its conceptual root in the idea of active interference (experiment, manipulability) with the 'normal' course of nature" (von Wright 1989, p. 833). The analysis of causation in von Wright's theory thus ultimately occurs at a very different level than in the Humean or Millian regularity analyses, and it does indeed rely on actions and agency. This is what I have chosen to call the conceptual part of von Wright's theory, and it is the significant part for our understanding the role of manipulations in his theory of causation.

In extreme brevity, von Wright's general idea is that human agency implies that we have the capacity to sometimes make certain things happen in the world that would otherwise not have occurred, and that we also then have the ability to refrain from performing such an action, and instead "let nature take its course." This introduces a branching of possibilities as to future events, that von Wright takes to be quite independent of the question of physical determinism. Likewise in our past we have had opportunities to act in one of several possible ways, and our choices of how to act or not act at all on those occasions determined the actual developments in the world. Thus our agency introduces *mere possibilia* into the ontology of facts and events, and von Wright claims that our belief in our own agency compels belief in this branching of world possibilities. It is this picture of a history of possibilities that accommodate an interpretation of the counterfactual conditionals that von Wright take as essentially a part of the relation of nomic dependence. (But he provides no semantic theory of counterfactual conditionals *per se* in this material.) Once histories of possibilities have been introduced, von Wright can identify causal relations within such histories, based on the conditionship relations that are present there.

In addition to the analysis and definition of causation, von Wright also treats the epistemic question of how we can and do *find* causal relations in nature, and this too can be viewed as a somewhat separate issue. Von Wright's suggestion on this matter depends as well on the particular role and properties he assigns to actions and manipulations.

4.2. Action and nomicity

One thing I believed that we could learn from Douglas Gasking's proposal was that for manipulations to do some specific work in a theory of causation, they need to be sufficiently specific things, with certain specific properties. We are interested in particular in how appealing to manipulations enables the analysis to distinguish cases of a causal relation, from those where a co-occurrence is due rather to a common cause, and to establish the direction of the causal relation, while at the same time managing to avoid a vicious circularity. Thus, questions about manipulations that we will want to ask, in light of the problems that they are intended to resolve, include "Do manipulations have causes (and if so of what kind)?", and "Do manipulations in turn cause their immediate results?" Von Wright's theory of causation comes out of his work in the philosophy of action (see von Wright 1971, Preface), so we can expect answers to at least some of these questions. In fact, his theory of natural event causes can be seen as an extension of his hermeneutical, teleological theory of action, that projects this perspective onto the domain of event causation. (For a brief overview of von Wright's theory of action, see Tuomela 1982.) I won't give an introduction to the philosophy of action in these pages, nor review or criticize von Wright's theory of action. Rather, I will treat the theory summarized in this chapter as a theory of causation proper, and in particular focus on what properties actions and manipulations have in it, that are relevant to our understanding of von Wright's manipulationist account of causation. That said, the most significant contemporary competitor to the sort of non-causal theory of action von Wright advocated was Donald Davidson's causal theory, introduced in "Actions, Reasons, and Causes" (Davidson 1963). According to this theory what makes some behavior (bodily movement) intentional, and thereby an action, is that it is *caused* by some intention of the agent whose bodily movement it is. Von Wright denied that actions could be distinguished from "mere behavior" (bodily movement without intention or purpose, such as the involuntary twitching of an eye or a stumble) by appeal to causation, in this way. The non-causal view of action was not exceptional at this time, but had rather been the dominant view during the 50's and early 60's (Stoutland 1982). Another alternative type of investigation into intentional behaviors was the functionalist approaches of for example Hilary Putnam (Putnam 1975) and Daniel Dennett (e.g. Dennett 1986), that identified intentional attitudes in an agent with certain functions, or roles, in the agent system. On these approaches, the understanding of intentionality is secondary, or subservient, to a mechanistic and physicalist view of events in general, and this is in contradiction with von Wright's position that agency and intentionality is at least as fundamental in the world as causation among natural events. Finally, von

Wright's account is also in disagreement with the agent causation theory of Roderick M. Chisholm, which proposed that the cause of an action is just the agent herself (Chisholm 1966), as we will see below.

Von Wright did not think, then, that causation could be employed to aid our understanding of action and agency, but that the reverse was true: to understand causation, we must consider its relation to agency, where the latter is considered as more fundamental, in some sense, than causation.

> My argument, to put it in a nutshell, will be this: The idea that causal connections are necessary connections in nature is rooted in the idea that there are agents who can interfere with the natural course of events. The concept of causation under investigation is therefore secondary to the concept of human action. (von Wright 1974, p. 1-2.)

Von Wright, then, aims to explain the concepts of causation and nomicity, or law-governed change in nature, by appeal to our beliefs about ourselves as agents with the capacity to perform actions. The core of von Wright's manipulationist analysis of causation is in fact—from the point of view of the metaphysics of laws of nature—a proposed resolution to the problem of theoretically distinguishing nomic from accidental regularities. From a more general metaphysical point of view, his appeal to agency is intended to resolve the familiar and controversial philosophical question of what a necessary relation, that is not *logically* necessary, could be (von Wright 1974, p. 9). It seems right to say, therefore, that this part of von Wright's theory of causation corresponds, in the larger scheme of theories, to David Lewis's "best system" theory of laws of nature, in the sense that it is what explains or grounds the meaning of certain claims about nomic modalities, which are subsequently used in the logical analysis of the causal relation (e.g.: Lewis 1973, sect. 3.3; Lewis 1994). This core, then, is von Wright's theory of nomicity and causal possibility that, once in place, allows for a process of "causal analysis" (von Wright 1971, p. 55) by which we can identify causal relations based on nomic dependencies, in a way largely familiar from Mill's analysis.

According to von Wright, when we truly believe that we have the ability to perform an action, such as opening a window, two things hold. Firstly, we can do it (open the window), and if we do it, then that the window opens is *not* an *effect* of our action. In a frequently quoted passage:

> The connection between an action and its result is intrinsic, logical and not causal (extrinsic). If the result does not materialize, the action simply has not been performed. The result is an essential "part" of the action. It is a *bad* mistake to think of the act(ion) itself as a cause of its result. (von Wright 1971, p. 67-68.)

In short, if the window did not open, then we did not perform the action of opening the window, even if we tried. The external event that has this intrinsic connection to an action von Wright calls its "result." When we "do p," this is an action that results in p. (If we take von Wright to mean here that the action cannot be the cause of its result because action and result are not independent

in the way required of cause and effect, we then want to note that Davidson would dispute this, claiming that only under *one*, but not the only possible, description of the action event, is the result logically implied (Davidson 1967).)

Many things we do, we do by doing something else. We may air the room by opening the window, and open the window by turning its handle. But ultimately there is something we do without doing something else, and these are intentional bodily movements. We may turn the handle by holding it and turning our hand, but we do not turn our hand by doing something else, and in particular we do not do it by sending the right signals along the nerves from our brain to our arm. Von Wright calls such actions that are not performed by way of doing something else "basic." The results of actions, then, are not *caused* by our performance of the action. This is not to deny that even basic actions are physical events. It is rather to acknowledge that some physical events are interpreted by us as intentional, and what this implies.

> A, the action, is M, the bodily movement, viewed (conceived, understood) under the aspect of intentionality. Viewing M under this aspect means relating it to the mental things R we call reasons for an action. This relation is not causal—although the fact that the reasons antedate the movement may create an appearance to the contrary. (von Wright 1998, p. 142.)

And: "[T]he determinants of action, I would maintain, are of a totally different kind from causes and effects among events in nature" (von Wright 1974, p. 2). Von Wright recognizes that the same bodily movement viewed instead under the aspect of a physical event can have ordinary event causes, and that under this aspect, what we correctly may call the result of the action, has that bodily movement as its cause (e.g.: von Wright 1971, p. 129; von Wright 1989, p. 806). We shall return in a later section of this chapter to the implications of this for the theory.

The second thing that holds whenever we truly believe that we can perform an action is that if we do *not* perform it, the thing we would have done will not happen anyway. This is a second way in which an action can fail: if the window *had* in fact opened, at that very moment, even had we not tried to open it, then we did not in fact open it when we tried to, and thought we did. Belief in agency implies, in this way, belief in certain counterfactual conditionals. Von Wright claims that these counterfactuals are not themselves causal (von Wright 1989, p. 830). Correspondingly, von Wright relates our general beliefs in ourselves as agents to certain facts about regularities in nature, that must hold if those beliefs are true. Firstly, it must be the case that often when we try to do p, p is subsequently realized. Secondly, it must be the case that when we believe ourselves able to do p but choose not to, p does not usually materialize anyway. If the first requirement does not hold, we lack the ability to do p after all. If the second requirement doesn't hold, we lack the *opportunity* to do p.

Now, if whenever we do p under a certain set of circumstances, q follows, and when we refrain from doing p under those same circumstances, q is also absent, von Wright takes it that we can *cause* q by doing p (and also prevent

q by refraining from doing p), and thus that p is a cause of q, under those circumstances. When von Wright claims that causation is "ontically" independent of agency, this is the reason: the causal relation holds between the thing done (p) and some other thing (q), both of which are natural events, and that relation exists whether p is the result of an action or comes about in some other way—causation is a relation between physical events in nature. (In contrast, the causal relation does *not* hold between the *doing* and the thing done.) The further thing q that p causes, when p is the result of an action, von Wright calls a *consequence* of the action (as opposed to its result p). Hence, on von Wright's theory, the causal relation is to be found as a pattern of co-occurrence among natural events, both actual and merely possible. It thus implies, and therefore relies on, certain counterfactual conditionals, and von Wright means that it is our convictions about our own capacities as agents that provide these counterfactuals, at the conceptual level. I'll now turn to his account of causal systems, or histories, wherein these counterfactuals have their interpretation, and in which causal relations can be identified.

The details of von Wright's account of nomicity and causal possibility goes through several apparent variations in the texts. In *Explanation and Understanding*, von Wright presents a graph of the kind shown in figure 4.3.1 on page 47 (von Wright 1971, ch. II). I'll call this a W-graph. This is a representation of a history of possibilities for a fragment of the world—what von Wright there calls a system. Each node in the graph denotes a partial generic world state, and each edge a possible transition of the system's state from one moment to the next. States are thus *types* on the model, that can have multiple token instances at different times. Branches toward the right in the topological tree thus correspond to the multiple possible future developments in the system. States in turn are composed of logically independent, simple generic states of affairs. Two nodes in the graph may denote the same state, the only limitation being that two nodes connected to the *same* immediately preceding node are taken to denote different states, to avoid redundancy. Causal relations are now identified as patterns of states of affairs in this structure, in a familiar way. That is to say, if a certain state of affairs p (which may be a part of several states denoted by nodes in the graph) is invariably followed by a state of affairs q everywhere in the structure, then p is causally sufficient for q, and if q is invariably preceded by p, then p is causally necessary for q. Taking the patterns in the structure to be nomic, this procedure is familiar from the traditional analysis of causal relations in terms of nomic dependencies, as exemplified by Mill and followers. Proper nomic dependencies are now distinguished from accidental co-occurrences of p and q by the presence of multiple branches, such that p is part of a node at a certain stage of the system's history on one branch, and absent from that same stage on a different branch, representing a different possible development of the system. The co-occurrence is nomic, or causal, if the regularity holds in all branches. In the absence of such branches, that p invariably precedes q would not be sufficient for saying that p is a necessary cause of q, since they may co-occur rather due to having a common cause.

This causal analysis depends, therefore, on the existence of multiple and unrealized possibilities. It is, as detailed above, to explain and justify the existence of such things—or perhaps just our belief in such things—that von Wright employs actions and agency.

In *Explanation and Understanding*, von Wright treats the topmost path through the graph as the *actual* history of the system. "Under this 'surface of reality' are the 'depths of alternative possibilities'" (von Wright 1971, p. 48). In "On the Logic and Epistemology of the Causal Relation" von Wright instead states that the topmost path in the graph represents the development of the system "'unless interference with the course of nature takes place." "What the topmost branches picture is thus the course of future developments if nature is 'left alone', 'untouched', 'to itself', to continue its course from any given point" (von Wright 1973, p. 303). Finally, in *Causality and Determinism*, von Wright gives no special interpretation to the topmost path of the graph at all. There is thus no representation in the theory of *actual* future developments, and von Wright's formal language has the corresponding inability to distinguish future actualities from other future possibilities. (This was observed by Prawitz in Prawitz 1989, p. 431.) Here, if there are multiple paths extending into the future, there may be no specific path that a system will take if left to its own devices—that is to say, the theory allows for physical indeterminism. However, what matters to von Wright's theory of nomic dependence is that actions may result in events that would not have occurred in the system on its "natural" trajectory. Agents thus affecting the system—this fragment of the world—are implicitly treated as existing "outside," or *exogenously* to, the system. This is not just a matter of how we commonly regard causal systems that we are interested in or, perhaps more illuminatingly, how we tend to regard *ourselves* in relation to such systems. As we have seen, von Wright takes actions to not have causes, and it follows that their influences must be exogenous to any causal system. In *Causality and Determinism* von Wright uses a +-sign inside a node to indicate when the possibility represented by that node depends for its realization on the action by some agent (von Wright 1974, p. 94). Thus, on this picture a causal system may be such that, on its own, it has a future trajectory that is perfectly determined by its prior states, but an agent can take it out of this predetermined path. In fact, von Wright goes as far as saying that some event may be a *physical* certainty, but its absence nevertheless a *causal* possibility, and that this depends on the presence of agency in the world. "If something is possible in a world without agency, it is causally possible. But not necessarily *vice versa*. Perhaps this thing can occur only with the 'help' of man" (von Wright 1974, p. 90). And: "The 'residue' of possibility that is not 'anihilated' by the physical certainty in question is the possibility that human action may make actual" (von Wright 1974, p. 92-93).

Thus, our picture of the world as one with an at least partially open future containing multiple incompatible possibilities, depends on the one hand on our belief that the world will develop in one way if we do not interfere with it, and on the other hand on our experiences of what happens when we *do* interfere. As it is inherent in the idea of agency that we may act to bring about

certain events, but also *refrain* from acting and let nature take its course, we must regard the future path of the world as contingent in these respects, and its history to consist partly of "lost possibilities." This results in histories of multiple possibilities, such that they can sometimes serve the purpose of causal analysis.

Of course, this is not the *epistemic* story, explaining how we can *discover* causal relations in nature. It is "only" the story that introduces the multiple future possibilities that accommodate meaningful causal counterfactuals in the first place. However, what is the intended status of this theory? Is this an explanation of human *beliefs* about causation, or about the *existence* of causal relations, or something else altogether? Von Wright is careful to point out that "[t]he existence of specific causal relations, and the operation of causal factors, is [...] independent of agency and the interference of agents with nature" (von Wright 1974, p. 49), and I gave his motivation above. But he takes a causal relation to be *conceptually* related to certain counterfactual conditionals (this part is largely uncontroversial), and these counterfactual conditionals in turn to be conceptually related to our notion of agency.

> It is true that, in addition to affirming a regularity, the causal nomic statement only says that, had the first of two states obtained when in fact it did not, then the second of them would on those occasions have obtained too. But in making this addition we employ a notion, *viz.* that of the counterfactual conditional which we should not have if we did not also have the notions of action and agency. (von Wright 1974, p. 53).

Von Wright's claim, then, is that causation is *not* ontically dependent on agency, but that there *is* a conceptual dependence. He also takes this as the reason why it is wrong to consider the relation between causation and agency to be purely epistemic. However, the obviously realistic mode with which von Wright introduces agency as a source of possible change in the world, that is independent of physical possibility, also suggests that the conceptual connection between agency and causation is not merely a question of the historical source of our causal concepts. This is also strongly suggested by his view of agency itself: "Perhaps we should call the fact that men can perform actions a 'mystery'—in the sense that it is something basic which defies explanation" (von Wright 1989, p. 809).

4.3. Defining causation

The previous section explained how von Wright aimed to ground a notion of nomic or causal necessity in facts about action and agency and certain kinds of counterfactual conditionals implied by these concepts. He employs these modal concepts in his analysis of the logic of the causal relation. Strictly speaking, von Wright differentiates between *causal analysis* and *causal explanation* (von Wright 1971, p. 55). In short, causal analysis identifies causal relations between generic events (types), while causal explanation gives part of the actual causal history of a particular (token) event, in terms of the types of the events involved and their type-level causal relations. Our focus will be on causal analysis, and

specifically the conditions von Wright impose on a type-level or general causal relation. Von Wright explicates these in a formal language, that is interpreted on histories of possibilities represented by W-graphs, such as the one shown in figure 4.3.1. We explained what these are in the previous section. (Von Wright also provides an axiomatization of this language, but I will only present its semantics here.)

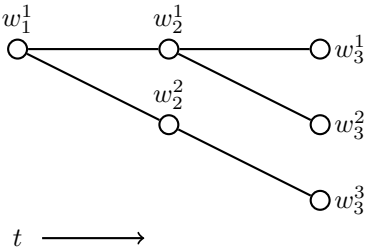

FIGURE 4.3.1. A W-graph. (Adapted from von Wright 1971, p. 50.)

This language has two kinds of modality: that of nomic or causal necessity, and that of temporal order. Moreover, it is a tense logic, meaning that one node in the model is singled out as the *now* of the system, and the truth of a proposition that employs the temporal operators may depend on what node this is. It is thus a tense logic of the kind mainly associated with Arthur Prior (Prior 1967), and expresses an A-theory of time. However, since what node in the model is the *now* does not make any difference to the truth of claims about general causal relations, in von Wright's interpretation of these, we will for simplicity assume that the *now* in the model is always the unique leftmost node.

In the definitions in table 4.3.1, that p obtains at a node w_j^i in a W-graph means that the state of affairs p is a "conjunctive part" of the state represented by w_j^i. In other respects, the language is an extension of a classical propositional logic. I have changed a couple of von Wright's symbols for typesetting reasons, but the wedges have their usual logical connotations, while an arrow here suggests a direction on the timeline (which is oriented the same as in the W-graph).

That von Wright's formal language is tensed is related to his view of the asymmetry of histories, that we described in the previous section: he takes the past to be fixed and linear, and the future to be open and branching. This has some affinity with the *growing block universe* theory of time, but is not exactly the same, since von Wright appears to assert that the future does exist, if only as branches of possible world developments. (Maybe we could call it the "Zipper Theory" of time—where the slider of the zipper is the moving *now* that transforms the open branches of the future into a closed and linear past.) Von Wright's reasons for thinking that the future is open and the past closed are not epistemic—he takes this asymmetry to be a real feature of time:

Operators for "in the future" and "in the past"

(i)	$\overrightarrow{\bigvee}p$	*iff*	p obtains at *some* node, in *some* future world branch.	
(ii)	$\overrightarrow{\bigwedge}p$	*iff*	$\neg\overrightarrow{\bigvee}\neg p$	
(iii)	$\overleftarrow{\bigvee}p$	*iff*	p obtains *now*, or at *some* past node.	
(iv)	$\overleftarrow{\bigwedge}p$	*iff*	$\neg\overleftarrow{\bigvee}\neg p$	
(v)	$\bigvee p$	*iff*	$\overleftarrow{\bigvee}p \vee \overrightarrow{\bigvee}p$	
(vi)	$\bigwedge p$	*iff*	$\overleftarrow{\bigwedge}p \wedge \overrightarrow{\bigwedge}p$	(p is "universally obtaining.")
(vii)	$\triangle p$	*iff*	$\overleftarrow{\bigwedge}p \wedge \overleftarrow{\bigwedge}\overrightarrow{\bigwedge}p$	(p is a nomic/causal necessity.)
(viii)	$\triangledown p$	*iff*	$\neg\triangle\neg p$	(p is a nomic/causal possibility.)

Operators for "right after" and "right before"

(ix)	$\overrightarrow{M}p$	*iff*	p obtains at *some* immediately succeeding node.
(x)	$\overrightarrow{N}p$	*iff*	$\neg\overrightarrow{M}\neg p$
(xi)	$\overleftarrow{M}p$	*iff*	p obtains at an immediately preceding node.
(xii)	$\overleftarrow{N}p$	*iff*	$\neg\overleftarrow{M}\neg p$

TABLE 4.3.1. Von Wright's logic of tense and causal necessity

> The openness of the future is here conceived of as an *ontic* openness. Ontically, the future is open (branching) and the past closed (linear). Past and future are thus, in a characteristic sense, *asymmetrical*. (von Wright 1974, p. 34)

This corresponds to the fact that there are no branches facing left, or toward the past, in W-graphs. It is also, of course, related to von Wright's explanation of causal possibility in terms of possible action by agents. But it does complicate his analysis of the causal relation, and in some respects—it would seem—unnecessarily. Note the following two peculiarities in the definitions in table 4.3.1 (both recognized by von Wright). Definition (vi) is the closest von Wright comes to expressing a universal regularity. But it's not quite that, because, as noted in the previous section, the language cannot express that something is a contingent future actuality. Hence, "$\bigwedge p$" means that p obtains, and has actually obtained throughout history, and will *necessarily* obtain in the future. Despite this difference, von Wright calls p when "$\bigwedge p$" holds but not "$\triangle p$" an "accidentally universal" state (von Wright 1974, p. 27). Secondly, as seen in

the table, "$\overrightarrow{M}p$" means that p is a possibility in the immediate future, and "$\overrightarrow{N}p$" that p is a necessity in the immediate future. But since a W-graph is always linear in the direction of the past, "$\overleftarrow{M}p$" and "$\overleftarrow{N}p$" mean exactly the same thing. As there is only a single node immediately accessible to the left of *now* in the graph, due to the linearity of the past, both "$\overleftarrow{M}p$" and "$\overleftarrow{N}p$" mean that p *actually* obtained in the immediate past. Von Wright calls the past branches in a W-graph that are not on its actual historical path the "lost possibilities." They are however not lost in the sense that they are unavailable to von Wright's logical analysis of the causal relation: they can be quantified over, as seen in definition (vii) of causal necessity. This definition states that p is a causal necessity in this system ("$\triangle p$") if and only if p always obtained in the past, and at every stage in the past, it was a future necessity. Or, equivalently in a non-tensed language: p obtains in all possible states of the system. Just as in more common analyses, it is the case that $\triangle p \Rightarrow \bigwedge p$ but not *vice versa*. We noted above that a W-graph represents a *fragment* of the world. It's easiest to understand what a fragment is by contrasting it to the history of possibilities for the whole world. A system then contains some continuous part of the world timeline, and some subset of the world's states of affairs, and these are the two ways in which the system can be a fragment relative to the world as a whole. The part of the world that is not explicitly modeled is often called the *environment* of the system. Clearly, if a state of affairs p is a causal necessity in the world as a whole, it is also a necessity in any fragment of it. But if p obtains in all possible states in a *system*, it doesn't follow that it is a causal necessity in the world as a whole. (The world may have possible states in which p does not obtain, but that are not part of the modeled system.) Therefore, to conclude that p is a causal necessity from the fact that it holds in every possible state of the system, we must assume that it either also holds in every possible state of the world as a whole or, more realistically, that it holds in every possible state of the world under certain background conditions, that are not explicit parts of the modeled system. Von Wright calls these relevant background conditions the *frame* of the system and, importantly, appears to identify the frame as part of the system (von Wright 1974, sect. III.5). That is, a different frame implies a different system. We can understand a system, then, as the time interval and set of states of affairs that are modeled, plus whatever part of the environment the dependencies between these modeled states of affairs in turn depend on.

"$\triangle p$" now expresses that p is a causal necessity. In a "first approximation" von Wright associates causes and effects with nomic dependency relations in a system: "$\triangle(p \rightarrow q)$" states that p is a causally sufficient condition for q, and also that q is a causally necessary condition for p. While von Wright recognizes that there are both causally necessary and causally sufficient conditions, and associates what we usually call a cause with a causally sufficient condition, he thinks that "$\triangle(p \rightarrow q)$" is problematically ambiguous, due to the asymmetry of the causal relation. It cannot be, on his view, that on one and the same occasion, p is causally sufficient for q and q causally necessary for p. Von Wright also admits of "negative" states of affairs, making available the contraposition of

the causally necessary implication, and further confounding what is said there to be causally sufficient for what. As in other treatments of the problem of the causal asymmetry, von Wright recognizes that under the assumption that causes always temporally precede their effects, this ambiguity is eliminated: "$\triangle(p \rightarrow \vec{N}q)$" expresses that p is immediately followed by q, by causal necessity, and it does not imply that $\triangle(\neg q \rightarrow \vec{N}\neg p)$. We can also say that an effect follows its cause after a certain number of state transitions, by iterating the $\vec{N}$-operator. Von Wright does however not accept the time-ordering analysis of the asymmetry of causation, although he thinks that such a time-ordering is ordinarily the case. We shall return below to his suggestion as regard this problem, which relies on possible manipulation.

The first approximation is further developed by stating that causation is a relation between *events*, not states of affairs. He defines an event as a change in what states of affairs obtain between two occasions in the system ("$\neg$" has been substituted for "$\sim$" in the quote, for consistency):

> [A]n event can be regarded as a change, or transformation (in time), among states. For example: the state p obtains on some occasion but $\neg p$ on a subsequent one. Then the event of p's passing away or vanishing or becoming destroyed has taken place. (von Wright 1974, p. 14.)

The *negation* (non-occurrence) of an event, in turn, is just the absence of a change to some state of affairs: either p obtains over two adjacent states, or it is absent in both (von Wright 1974, p. 72). To save some space, I will by a "p-event" mean an occurrence where a system passes from a state where p does not obtain to one where it does. Following conventions elsewhere in this text, I denote by "$\mathtt{p}$" a particular p-event. Since causes and effects as now imagined are things that exist across two subsequent states, rather than *at* a state, von Wright provides a new formulation of a causal law (I will call it *EC* for "event causation"):

$$EC\text{:}\ \triangle(\neg p \rightarrow \vec{N}(p \wedge \neg q \rightarrow \vec{N}q))$$

EC states that "It was and will always be the case that if p was or would have been absent from the world on some occasion, then it was or would have been certain that either it is *not* the case that p is present and q absent on the immediately succeeding occasions *or* it is certain that q is present on the occasion immediately after that one" (von Wright 1974, p. 83-84). A positive instance of this law would be expressed "$q \wedge \overleftarrow{M}(p \wedge \neg q \wedge \overleftarrow{M}\neg p)$." As expected, EC implies a mere regularity (or what passes for one in von Wright's theory) $\bigwedge(\neg p \rightarrow \vec{N}(p \wedge \neg q \rightarrow \vec{N}q))$. As the time difference between cause and effect may vary, this is just one example of a causal law. In particular, judging from what von Wright had said in earlier sections of *Causality and Determinism* (e.g. von Wright 1974, sect III.1) we must assume that there is an expression of a law of simultaneous event causation, although von Wright does not provide

such an example. A candidate may be "$\triangle(\neg p \wedge \neg q \rightarrow \vec{N}(p \rightarrow q))$." Given the problem of the causal asymmetry that von Wright discusses, we would expect this law of simultaneous causation to be unclear as to what is cause and what is effect. To illustrate, if we contrapose the second implication—resulting in "$\triangle(\neg p \wedge \neg q \rightarrow \vec{N}(\neg q \rightarrow \neg p))$"—we go from a statement about a change in p being sufficient for a change in q to one of a non-change in q being sufficient for a non-change in p. But if the occurrence of a change in p is accepted as a cause of a change in q, we are not therefore committed to a non-change in q, under those same circumstances, being a cause of a non-change in p. In an example where there is a temporal ordering, rain may cause the ground to get wet, but dry ground is not a cause of the absence of rain. In cases where we cannot distinguish cause from effect based on a time-ordering, some further condition is thus required.

Von Wright notes that there can be p and q, and histories, such that EC is satisfied, but where we would not recognize this as a causal law. He describes three cases of "trivialization": *(i)* a change from $\neg p$ to p is causally impossible; *(ii)* a change from $\neg q$ to q is causally necessary; *(iii)* $p \wedge \neg q$ is causally impossible after $\neg p$. Two possibility conditions eliminate these cases. For *(i)* and *(iii)*, $\bigtriangledown(\neg p \wedge \vec{M}(p \wedge \neg q))$ and for *(ii)* $\bigtriangledown(\neg q \wedge \vec{M}\neg q)$. We can then give the full expression of the causal law:

$$EC'\textbf{:}\ \triangle(\neg p \rightarrow \vec{N}(p \wedge \neg q \rightarrow \vec{N}q)) \wedge \bigtriangledown(\neg p \wedge \vec{M}(p \wedge \neg q)) \wedge \bigtriangledown(\neg q \wedge \vec{M}\neg q)$$

Since von Wright essentially distinguishes nomic from accidental regularities by the fact that the former imply certain counterfactual conditionals while the latter don't, it is important that this is a result also in the expression of causal laws in his formal language—and it is. His counterfactual conditionals are interpreted differently from how it is done in the Lewisian tradition, that has since become common. Von Wright also only provides an expression for the counterfactual conditional implied by his "first approximation" causal law, and not the causal law between events expressed in EC'. Thus, the counterfactual conditional that follows from "$\triangle(p \rightarrow \vec{N}q)$" is "$\vec{\bigwedge}\overleftarrow{\bigwedge}(\neg p \wedge \overleftarrow{M}\vec{M}p \wedge \overleftarrow{M}\vec{M}\vec{M}\neg q \rightarrow \overleftarrow{M}\vec{N}(\neg p \vee \vec{N}q))$." This expression is read as "It will always be the case that, it always was the case that, if p is absent, and p could have been present, and it could have been the case that q will be absent in the next moment (not taking into account that p is absent), then it is necessarily the case that either p is absent or q will necessarily be present in the next moment." This follows from the approximate law because the consequent of the conditional follows from it. (I.e., for an arbitrary *now*, $\triangle(p \rightarrow \vec{N}q) \Rightarrow \overleftarrow{\bigwedge}\vec{\bigwedge}(\neg p \vee \vec{N}q) \Rightarrow \overleftarrow{M}\vec{N}(\neg p \vee \vec{N}q)$.) This counterfactual conditional does not follow from a mere accidental regularity $\bigwedge(p \rightarrow \vec{N}q)$, since even if the material implication always held in the past, it doesn't follow that its negation was never a possibility, in one of the "lost branches" of the system.

In his discussion about how we can discover causal relations by empirical investigation von Wright introduces one more qualification as to what causal laws are, and we will return to this in the section on the epistemics of causation, below.

This part of von Wright's theory, his causal analysis, does not mention manipulations at all. We may now return to the quoted definition of causation with which we began this chapter.

> 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.)

To understand this definition better, we should summarize what we have said about von Wright's theory so far. Von Wright takes general causal claims to state a causal law. Such laws are distinguished from merely accidental co-occurrences in that they imply certain counterfactual conditionals: if a p-event is a cause of a q-event, then on an occasion when p did not obtain, if it *had* come into being, so would q. Von Wright thinks that such counterfactual conditionals could only be implications of our beliefs about ourselves as agents. That is, we may in this case *do* p, or refrain from doing p, and q would then come into being or not, respectively. So, when von Wright anticipates the objection that it is causation that makes manipulations possible, and not the other way around, he claims that such an argument would "beg the question" (von Wright 1971, p. 71). It simply fails to account for the counterfactual conditional that separates nomic or causal co-occurrences from accidental ones. To von Wright, this is the counterfactual implied by our ability to act (as described in section 4.2), and it is *not* causal. On von Wright's view, then, it must be that every cause is an event of a type that we may bring about as a result of an action—or it must be *composed* of such things.

With respect to this extension of the definition to events that we clearly could not bring about, von Wright takes the eruption of Vesuvius as an example. This eruption was the cause of the destruction of the city of Pompeii, and it could not have been brought about by man. But

> [w]ithin each of [these events] a number of events or phases and causal connections between them may be distinguished. For example, that when a stone from high above hits a man on his head, it kills him. Or that the roof of a house will collapse under a given load. Or that a man cannot stand heat above a certain temperature. All these are causal connections with which we are familiar from experience and which are such that the cause-factor typically satisfies the requirement of manipulability. (von Wright 1971, p. 70.)

We might say, then, that von Wright introduces yet another way, as compared to Gasking's proposal, in which we can extend the class of causes from the class of manipulable events. An event is of the right kind for being a cause if it is of a type which has manipulable instances, but it is also of the right kind if it is a composition of such events.

On von Wright's view, the causal relation is, despite the essential role of agency in the account, a relation between natural events—if p-events cause q-events then they will do so whether they were brought about by an action or not—and the causal relation can therefore be said to be ontically independent of agency and manipulation.

We can return now to the questions we posed at the beginning of the chapter. We have reason to think, on von Wright's account, that manipulations are independent of the causes in the systems we manipulate, since actions do not themselves have causes. (We will however complicate this picture below.) Manipulations do not, moreover, cause their immediate results, on von Wright's theory, since the result is logically, and not causally, connected to the action. In short, if my action consists in that I open the window, it follows directly that the window opened. If it didn't, then I failed to perform that action. Hence, there can be no causal relation between my action of opening the window and the window opening; they are just not logically independent in the right way. These properties of manipulations are important for how manipulations can be used to resolve some metaphysical and epistemic problems that will remain in focus throughout the rest of this chapter.

4.4. The problems of nomic necessity and causal asymmetry

Before summarizing von Wright's account of how we can discover causal relations, I want to emphasize the way in which his theory engages with two traditional theoretical problems that we have already encountered, namely that of how to understand the necessary connection between events that has classically been associated with causation, and that of establishing the asymmetry of the causal relation.

In the first chapter we noted the skeptical, empiricist view of the idea of a natural necessity that distinguishes laws from mere accidental regularities, which derived mainly from Hume. Collingwood, too, rejected this sort of necessity. To him, the idea of causal necessity is derived from our experience of agency. Gasking, we might say, avoided the explicit issue of necessary connections and laws of nature by directly connecting causation to human ways of manipulating events in the world, and extrapolating from these to unmanipulable causes. Von Wright's theory has similarities to both. We have seen how he expands the domain of causes to unmanipulable events in a way similar to Gasking's. But he also says that, "the distinction between cause- and effect-factors [among natural events] goes back to the distinction between things done and things brought about through action" (von Wright 1971, p. 73), and that if we produce p and thus bring about q, then "we vest the first of the two states [p] with a 'power' of producing the second state [q], analogous to our power of producing the first" (von Wright 1974, p. 51). These are explanations of the conceptual connection between causation and agency in von Wright's view. Now, we can easily imagine that Collingwood might concede that this is what is happening, but that he would take it to be a *misapplication* of the concept, when it is applied to unmanipulable things, and that the talk of the first state's power to bring about the second state fails to correspond to anything

real. Since von Wright emphasizes repeatedly that the existence of causation is independent of agency and our ability to manipulate, he appears to embrace the reality of a necessary connection in causal laws (by way of their counterfactual implications), and his view is then more complex than Collingwood's. Thus, von Wright aims, as we have seen, to explain at least the conceptual distinction between causal laws and accidental regularities in terms of the counterfactual conditionals implied by the former but not the latter, and these counterfactuals in turn by appeal to action and agency.

Gasking's proposal was mainly meant to address the problem of the asymmetry of the causal relation, given that we reject the time-ordering condition. We saw that in von Wright's causal analysis, what is a cause of what was not determined by the dependency relations alone. The analysis could therefore not distinguish, in the expression of a causal law, between a change in p being a cause of a change in q on the one hand and a non-change in q being a cause of a non-change in p on the other, since the second causal sufficiency relation is just the logically implied contraposition of the first. Von Wright moreover rejected time-ordering as the grounds for the asymmetry of the causal relation. In fact, in *Explanation and Understanding* he embraces the actuality of time-reverse causation in some situations—specifically when raising one's arm is causally sufficient for an *earlier* neural event in one's brain (von Wright 1971, p. 76-77). (This kind of case had been previously discussed in Chisholm 1966). However, he later describes this particular section of *Explanation and Understanding* as "defective and unconvincing" and no longer accepts its argument for time-reverse causation (von Wright 1989, p. 811).

Von Wright appeals to possible manipulations directly to distinguish causes from effects among causally related events. In a particular kind of circumstance, either one of the events was *actually* manipulated, which settles the issue, or if no event was manipulated we can understand by analogy to situations with which we are familiar that one of the events occurred *as if* it had been manipulated. If we are unsure about the latter kind of case, we can often investigate it by performing an experiment, again relying on an actual manipulation to establish the direction of causation. Note that, for what has been said so far, the manipulated event *must* be the cause, due to the properties von Wright's theory attributes to manipulations: if a and b co-occur and the event a is a direct result of an action, then, since the action does not have a cause and a is its result, a does not have an external event cause, and in particular it does not have b as its cause. Nor does a have a cause in common with b. Hence, a must be the cause of b. We could say that this result is a consequence of the manipulation being taken in the theory to be causally exogenous to the manipulated system, in the strongest possible sense. This brings us to the epistemics of the causal relation.

4.5. Finding causal relations

Von Wright explains how we can discover causal relations between events in somewhat different ways in *Explanation and Understanding* (von Wright 1971, sect, II.7), "On the Logic and Epistemology of the Causal Relation" (von

Wright 1973), and *Causality and Determinism* (von Wright 1974, sect II.3 and III.7-9), but the underlying general principles are common throughout. In explaining the logic of causal discovery in von Wright's theory, and the role of manipulations in this, we should start with a reminder of what causal laws are. Von Wright takes a causal law to be a regularity in a history of possibilities, expressed by a W-graph. He calls such a history a "system" in *Explanation and Understanding*, and a system is a fragment of the history of a world. This regularity can be understood in terms of dependency conditions (what von Wright calls "conditionship relations"), such that a p-event is a cause of a q-event only if an occurrence of a p-event is sufficient for an occurrence of a q-event in the system. In addition to this, von Wright adds two possibility conditions to exclude cases where such a regularity holds trivially. Above, EC' was the expression of one example of a causal law in von Wright's formal language. In his discussion of the epistemics of causation, he seems to implicitly weaken this condition somewhat (von Wright 1974, sect. 8-9). There, he introduces a new notion of "physical necessity" which is a state of affairs that is certain to obtain *unless the system is interfered with* by an agent. He supplies a new way of expressing this: "$\vec{N}_\phi p$." Correspondingly in the W-graph, a node that can only be realized under a manipulation is marked by a "+." The general idea is now that manipulations somehow allow us to investigate different branches of a system, and in that way provide information about the dependency relations in it.

> Metaphorically speaking, what is required [. . .] is a dive under the surface of actual reality into the depths of unactualized possibilities, the "lost possibilities" of an ever growing past. (von Wright 1974, p. 37.)

Von Wright of course recognizes that this is not strictly speaking possible—we cannot observe counterfactual situations—but he thinks that we can come sufficiently close to form reasonable beliefs about causal dependencies in observed systems. It is clear from what von Wright says most of the time that he thinks of his systems as *tokens*, part of the history proper of the world, such that once they have occurred, they can never be observed again. But when he speaks in this way of empirically testing different possible evolutions of a system, it is also clear that one way of theoretically accommodating this is to regard systems as *types*, that can be instantiated on multiple occasions. Since the states in von Wright's systems are composed of generic states of affairs, the difference between the token view and the type view of the systems consists only in the fact that in the latter, the temporal relations within the system are relative, for example to the initial state, rather than mapped to objective points in time.

Figure 4.5.1 shows a system that, if left to its own devices, will go through three stages, in which $\neg p$ obtains during the first two stages, and $\neg q$ obtains during the last two stages. However, an interference by an agent may bring p about in the second stage. If the system is then left alone, it will enter a state where q obtains. But yet another interference may prevent q to come about in the last stage. This last possibility is what von Wright makes room for in his discussion about how we find causal relations, and that requires a

weakening in the expression of the causal law. We can express the corrected law:

$$EC''\text{:}\ \triangle(\neg p \rightarrow \vec{N}(p \wedge \neg q \rightarrow \vec{N}_{\phi} q)) \wedge \bigtriangledown(\neg p \wedge \vec{M}(p \wedge \neg q)) \wedge \bigtriangledown(\neg q \wedge \vec{M} \neg q)$$

Moreover, von Wright introduces a condition that must hold for it to be in general possible to perform a "test of nomicity" for some observed regularity. The condition, $\bigvee(\neg p \wedge \vec{M}(p \wedge \neg q) \wedge \vec{M}(\neg p \wedge \neg q) \wedge \vec{N}_{\phi} \neg p)$, states that $\neg p$ is possible, and after such a state, both $p \wedge \neg q$ and $\neg p \wedge \neg q$ are possibilities, and—crucially—$\neg p$ will certainly follow the initial condition unless a manipulation brings p about.

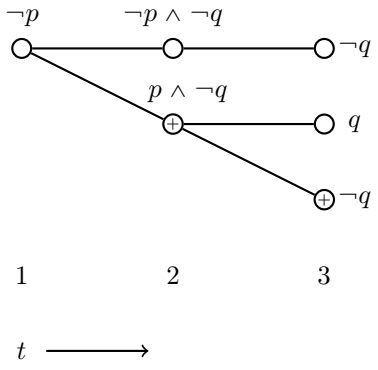

FIGURE 4.5.1. A W-graph satisfying EC''.

Figure 4.5.1, then, shows a system that satisfies EC'' as well as the condition for a possible test of nomicity. In reconstructing, somewhat, von Wright's reasoning, we might now want to consider what exactly it is we can observe. By passively watching this system develop, we would at the third stage observe that $\neg q \wedge \overleftarrow{M}(\neg p \wedge \neg q \wedge \overleftarrow{M} \neg p)$. When we bring p about through a manipulation in the second stage, and after that let the system evolve without interference, we would at the third stage observe that $q \wedge \overleftarrow{M}(p \wedge \neg q \wedge \overleftarrow{M} \neg p)$. We would thus have observed the two possible evolutionary paths of the system that establish that the regular co-occurrence of the coming about of p and the coming about of q is causal rather than accidental. Causal discovery depends, according to von Wright, on at least two observations of a system, one passive and one "active," in this way. Of course, our conclusion that this regularity is causal, based on these observations, depends on several strong assumptions, that may for all we know be false.

It is convenient in this context to adopt the assumption that W-graphs represent system types. Only then does it make straightforward sense to suggest that we can observe the realization of different evolutionary paths for the *same* stages of a system at all. Above, we noted that a system is not just individuated on what is explicitly modeled in its W-graph, but also on what von Wright calls its implicit *frame*. More generally this is what we have called

the causally relevant background conditions, under which a causal relation is assumed to hold. Now, from a very general point of view, the difficulty in assessing whether an observed regularity is causal, even under an experimental trial such as described here by the manipulation of the system, lies in knowing that we are really observing an instance of *the same* system in the cases of passive and active observation. That is to say, if the implicit frame of the system has changed in between those occasions, it is quite possible that a q-event would have occurred in the third stage of the manipulated system also had it *not* been manipulated, and our conclusion that it was the occurrence of a p-event that caused the occurrence of a q-event would then be false. This is the expression in von Wright's theory of a completely general problem with causal discovery—it's reliance on that the relevant background conditions remain the same between observed instances. (This is not a problem unique to von Wright's proposal, of course, but a perfectly general concern in inductive causal discovery.)

So it may well be the case that in a system S, a q-event never occurs at the third stage, whether we bring about p or not, and in a similar-looking system S', a q-event always occurs at the third stage, whether we bring about p or not, and that we are confusing these two systems in our observations. To see now why it is important in this theory that it is a *manipulation* that brings p about, observe how it excludes one natural way of understanding what is going on in such a confounding situation. Assume that we have observed a q-event at the third stage on all and only those occasions when a p-event occurred at the second stage. Then it might be that in S, no sufficient cause of a q-event at the third stage occurs (modeled or in the frame), so q never obtains there, in any possible system evolution. And in the observed instances of S', a *common cause* `r` of the p-event at the second stage and the q-event at the third stage is part of the frame, so q always obtains at the third stage. (If the p-event is a barometer dip and the q-event is rain, the r-event may be falling atmospheric pressure.) Thus we have *not* observed a cause `p` of q's coming about. However, if `p` really was brought about by an action, then, on von Wright's theory, what we just described cannot be the case. This is because if `p` was the effect of some natural event, then that event may also be the cause of `q`, but if `p` was brought about by an action, then it has no cause at all. As explained at the end of the last section, this means that `p` cannot have had a shared a cause with `q`. (Nor can `q` have been the cause of `p`.) Thus, under the assumption that the observed co-occurrence of p-events and q-events is not a coincidence, the only remaining explanation is that p-events cause q-events.

> Would not the mere occurrence of E, followed by F, be just as relevant to our causal hypothesis as the production of E? The answer is no. For in the case of a 'mere' occurrence we should have to reckon with the possibility that E was in its turn caused by something that also causes F. And then E need not be a cause of F at all. If, however, E did not merely occur but was produced by an action, no such common cause of E and F can be there. Its existence would contradict the

> assumption that E resulted from an interference. (von Wright 1989, p. 831.)

This is how von Wright's theory of actions operates in the context of the logic of causal discovery. It has the theoretical force it has by making actions causally exogenous, and thus independent, of any natural systems of events whatever.

Note also how the sort of situation described by the first stages in figure 4.5.1 is precisely what von Wright takes to be *implied* by our belief in our own capacity as agents, relative to some p-event. He had stated that to believe that one can *do* p implies two things: first, that if we refrain from doing p, p will not come about anyway, and second, if we try to do p, often enough we will succeed. This maps neatly over the W-graph in figure 4.5.1, and illustrates how our beliefs about causal relations are connected to our beliefs about our own agential capacities.

4.6. Actions and bodily movements

Let's reiterate the argument that, firstly, presents this manipulationist solution to the problem of analytically distinguishing a co-occurrence that is due to a direct causal relation from one that is due to a common cause, secondly, establishes the direction of the causal relation, and finally thereby shows what is a cause of what in a causal system. First, since actions do not have causes, their results do not have a cause, and therefore cannot share a common cause with a correlated remote (relative to the action) event. Second, for the same reason it cannot have that remote event as its cause. Thus, a regular co-occurrence of a p-event and a q-event that is present as p is manipulated must be due to the p-event being a cause of the q-event (ignoring the possibility that this regular co-occurrence is mere chance). Part of my purpose in putting the reasoning in this way here is to connect von Wright's use of manipulations to later accounts relying on *interventions*, defined more specifically in causal terms. More specifically, we can employ in our understanding of von Wright's theory what we have learned from those later theories about the conditions under which we can infer the presence of a causal relation. While von Wright never calls manipulations "causally exogenous" to manipulated systems—this is a later notion in the philosophy of causation—the gist of the argument I have formulated here is never far below the surface of von Wright's presentation. It seems to me to be the best—perhaps the only—way to try to defend the results claimed for the theory, even if that defense were to fail in the final analysis.

The claim that actions don't have causes, then, is instrumental to this reading of von Wright's account of causation, both in theoretically distinguishing the direction of causation among causally dependent events, and in the logic of causal discovery. This is just because it makes actions causally exogenous to any manipulated system. However, the claim that actions don't have causes can seem implausible or difficult to understand on the face of it, considering that actions are bodily movements, and these in turn physical events, some of the physical causes of which we moreover think we know a great deal about. Von Wright is in fact a compatibilist in this respect. We quoted him above saying that an action is a bodily movement understood under the "aspect of

intentionality." But the non-intentional aspect of the same bodily movement is also available, and von Wright thinks that "the interpretation of behavior as action is compatible with the behavior having a humean [event] cause" (von Wright 1971, p. 129). He also considers "psycho-physical parallelism"—the "correspondence between a causal and a rational explanation of a chain of events [some of which are bodily movements]"—to be plausible (von Wright 1998, p. 134). And finally: "I accept [...] what is called the Compatibility Thesis. A causal explanation of why my arm rises [mere behavior] is compatible with a teleological explanation of why I raise my arm [intentional act]" (von Wright 1989, p. 806).

But this presents a problem for his manipulationist theory of causation. Because if a bodily movement that we *may* understand as an action has an event cause under *some* way of understanding it, then that cause may after all be a common cause of the bodily movement and a remote event that correlates with that bodily movement, or it may be that very correlated remote event. In short, exogeneity is not longer guaranteed, and thus both the solution to the asymmetry problem and this way of finding causal relations through manipulations would appear to be blocked.

Von Wright discusses the relation between the physical and the intentional at length in "On Mind and Matter," included in (von Wright 1998). That discussion, however, does not provide any straightforward answers to the issue raised in the previous paragraph, and I don't know how von Wright would respond to it.

4.7. Reviews and criticisms

Although I believe that it is central for understanding von Wright's theory of causation at all to acknowledge how closely it connects to his theory of action, the debate in the philosophy of action proper is—as already noted—well beyond the scope of this chapter. Some philosophers who have engaged with von Wright's theory of action are Donald Davidson (Davidson 2001), Frederick Stoutland (Stoutland 1982, Stoutland 1989), and Alan Donagan (Donagan 1989). Here I will treat only reviews and critiques of von Wright's theory of causation, as compared to other theories of causation.

Many of the objections applicable to von Wright's theory have been directed at manipulation accounts of causation in general. I will focus on criticisms aimed explicitly at von Wright's theory when possible, and that were published before Peter Menzies and Hugh Price's updated manipulationist proposal, that we will spend the next chapter on (Menzies and Price 1993).

4.7.1. Circularity. We looked at Alexander Rosenberg's criticism of Douglas Gasking's manipulationist theory of causation in the previous chapter (Rosenberg 1973), and in particular his claim that Gasking's theory was viciously circular. That critique was actually aimed at von Wright's then recently published account, and prompted by von Wright naming Gasking's theory as the one most similar to his own (von Wright 1971, p. 189, note 40). Circularity is perhaps the most common objection to manipulationist theories in general (e.g.: Hausman 1997, p. S17; Paul and Hall 2013, p. 38; Psillos 2014, p. 103;

Woodward 2014a, p. 1715). To repeat, the problem has been taken to be that, since "to manipulate" is itself a causal notion, meaning something along the lines of "causing to change," appealing to manipulations in grounding or explaining causation involves causation in those grounds or explanations. Thus, the complaint also amounts to the theory not being a successful analysis of causation in other terms. Daniel Hausman has brought this criticism up repeatedly over the years (e.g.: Hausman and Woodward 1999; Hausman 2008, p. 89). In 1986 Hausman stated:

> The circularity [in the manipulationist theory] is evident: To affect a variable is to have a causal influence on it. If one does not already understand what causation is, and if one does not already know that one's interventions cause the alterations and are not caused by them, the manipulability theory will tell one nothing. (Hausman 1986, p. 145.)

As we noted in section 4.3, von Wright went out of his way to explain that actions do not cause their immediate results, but that those results were a logical part of the action. Hausman's response to this is that "one may always convert refutations into philosophical mysteries" (Hausman 1986, p. 145). But von Wright's argument seems essentially correct to me. The "mystery" appears in the next step of the analysis. Similarly as in the manipulation case, if the storm broke the window, then it follows that the window broke, and on von Wright's account there can be no causal relation between the storm's breaking the window and the window breaking, since the dependency relation between the two is not of the right sort. Nevertheless, no one doubts that if there is causation at all, "the storm broke the window" describes an instance of it, and this is not contradicted by the supposition that the event thus described cannot be the cause of the window breaking. The cause of that, of course, would then be the storm. Somewhat analogously, von Wright does acknowledge that when he opens the window, then the movement of his hand may be (i.e., it is not excluded by his theory) the cause of the window opening. But in the statement "I opened the window," von Wright likely does not think that "I" describes the cause of the window opening. "Action language" takes us into the intentional domain, where the concepts of event causation do not apply.

So, is von Wright's analysis of causation circular? At least not on the face of it. We need to remember that the real analysis is occurring, not directly by way of the definition of a causal relation in terms of manipulations, but at the level of causal counterfactuals and the distinction between accidental and causal regularities. I do agree that the fact that we can switch between the intentional aspect and the non-intentional aspect of certain events creates a real mystery in the account. But it would seem that von Wright is keen on acknowledging this as well, and even to suggest that this is a real mystery existing in the world.

4.7.2. Anthropomorphism. The charge of anthropomorphism amounts to the claim that manipulationist theories, by making the presence of causation dependent on a possible manipulation, also make causation dependent for its nature and existence on the particular faculties and capacities of human or

other agents. Hausman, again, states rhetorically, with respect to the asymmetry problem: "Are causal asymmetries dependent on human actions and perspectives, or are they objective facts?" (Hausman 1986, p. 145). Mackie, in turn, says in *The Cement of the Universe*:

> It certainly seems that our voluntary actions give us our primary, direct, awareness of causal priority. It is not unlikely that something that belongs essentially to them is at least the core of our concept of causal priority. Nevertheless, there may be something of which we are then at least dimly aware, something objective and not essentially tied to human agency and intervention.
>
> [...]
>
> Surely von Wright has mistaken the experience which gives us our primary awareness of causal priority for the relation of causal priority itself. (Mackie 1980, p. 171-172.)

Now, it is clear that von Wright's theory is *intended* to be agent-centric, it is the very philosophical outlook of his theory of action, of which his theory of causation is an outgrowth. This is true also for the theories of Collingwood and Gasking. I think there is a fairly straightforward historical explanation for this approach, which seems to fit Collingwood's and Gasking's cases well. As recounted in the first chapter, Bertrand Russell claimed in 1912 that the laws of nature are not causal, are not described as causal in the advanced sciences, and that the cause concept is empty and a source of philosophical confusion that needs to be eliminated. The approaches of Collingwood and Gasking can then be seen as defending the meaningfulness of the concept of cause in the realm of human activities and interests, without making it fundamental to the workings of the world—that is, as saving the concept, while acknowledging Russell's main point. This is clearly one of Collingwood's goals. While this may explain somewhat the occurrence of these theories at this particular time, and show why their agent-centrism is precisely the point and not obviously problematic, the situation is considerably complicated in von Wright's theory. Already between Collingwood and Gasking there was a move toward acknowledging objective causation between events. Von Wright now goes still further, by insisting that causation is *not* ontically dependent on manipulation—it is an objective relation between natural events that exists whether something is manipulated or not. Thus the possible connection to Collingwood's goal of saving the cause concept, while acknowledging Russell's criticism of it, is less clear. Again, however, von Wright's main manipulationist thesis is that the *concept* of a causal regularity is dependent on the concept of action. This conceptual connection is clearly not supposed to be merely a contingent *psychological* fact in von Wright's theory, about what we are able to conceptualize, and in virtue of what we have that ability. I'll highlight here the response that von Wright gave to the sort of objection expressed by Edwin McCann in his review of *Causality and Determinism*. McCann says that von Wright's way of accommodating the asymmetry in causation

> will be apt only if the "actionist" account of causality is convincing, for to those lacking such convictions it will seem more natural to suppose that our ability to "produce one state and thereby bring about another," rather than providing for the one state's being the cause for the other itself vitally depends on that fact. (McCann 1978, p. 89.)

In his anticipation of this objection, von Wright had responded (as McCann recognizes), that the objection fails to account, as it must, for the causal counterfactuals that distinguishes causal from accidental regularities, and that the only way to do this was by way of our beliefs about our own capacities as free agents. As indicated by this response, von Wright's theory is not at its core naturalistic (as for example Daniel Dennett's "intentional stance" theory is). We know that his compatibilism recognizes that there may be a true causal account even of the concept formation processes underlying our beliefs about ourselves as agents, but this account, too, depends on and is in some sense secondary to, those beliefs, which must be taken as basic.

So, is von Wright's theory anthropomorphic? Yes, and it's the central feature of the account. Thus, similarly to what I believed to be the case with Collingwood's theory, the anthropomorphism as such is not really a problem *in* the theory, as much as a problem *with* the theory, for those starting from more naturalistic presuppositions.

4.7.3. Mysterious modalities. Natural modalities have been regarded as unacceptably mysterious, among empiricists in particular, and are still regarded so, by many regularistically inclined philosophers. We find the view expressed in Frank Ramsey's "Law and Causality":

> But may there not be something which might be called real [necessary] connections of universals? I cannot deny it, for I can understand nothing by such a phrase; what we call causal laws I find to be nothing of the sort. (Ramsey 1978, p. 148.)

Ayer, in turn, spoke of the "mysterious property of being necessary" (1956). In "New Work for a Theory of Universals" David Lewis rejects Armstrong's necessitarian theory of laws, because he "find[s] its necessary connections unintelligible" (Armstrong 1983; Lewis 1983, p. 366). Norman Swartz said, with regard to the supposed difference between an actual regularity that is due to a law and one that is not, that "[a]ttributing the source of the nomicity of the one and of the accidentalness of the other to a physical possibility of 'the world as a whole' remains unintelligible" (Swartz 1985, p. 103). Here, physical possibility is of course the correlate of a necessary relation between particulars, that may be called "physical necessity." As a final, recent example, Jonathan Schaffer says that "[m]odal entities [such as a necessary relation between particulars] by themselves seem shadowy and mysterious. It seems they cannot float free – they need grounding in the occurrent." (Schaffer 2007, p. 85).

It may be unsurprising, therefore, that several commentators have found von Wright's account of causal and nomic modalities unsatisfactory. McCann

focuses on the fact that the action counterfactuals, that are explicitly non-causal and on which the causal counterfactuals conceptually depend, remain primitive and unanalyzed in von Wright's theory (McCann 1978, p. 92). This is how the theory is consciously constructed, and von Wright clearly takes these counterfactuals as the appropriate place to begin the analysis—McCann clearly disagrees. Again, the difference appears to be an expression of different philosophical expectations. It therefore seems hard to me to make any sort of universally convincing objection to von Wright's proposal along such lines. Von Wright may well be regarded has having given an adequate account—even exactly the right sort of account—of nomic necessity, to those philosophers who are willing to take the presence of free human agency as a fundamental fact about the world. A more universally successful objection would then need to show that von Wright's theory somehow fails in its stated goals.

Mackie gives a specific argument to the effect that von Wright fails to illuminate the nature of his nomic necessities after all. He notes that on von Wright's view the belief that we can *do p* under some circumstances entails two other beliefs, about regularities: that *p* will usually come about when we try to "do it," and that *p* usually will *not* come about when we do not try to do it. He then argues that the second belief, about a nomic regularity in the absence of interference, on its own has counterfactual implications. And that

> anything that gives us a good reason for believing that situations of a certain sort *will* remain stable (at least for a while) in the future equally gives us a good reason for believing that a similar situation, which we disturbed, *would have* remained stable (at least for a while) if we had not intervened. But, if counterfactuals can be supported in this way, it is not clear that we need human action to "verify" causal counterfactuals, and von Wright's case for the conceptual dependence of natural causation upon human action begins to falter. (Mackie 1976, p. 215-216.)

Mackie concludes that the objective facts that von Wright points to in his defense of the objective nature of causation are in the final analysis just these particular, actual regularities, and therefore that "von Wright's causation is Humean, all too Humean" (Mackie 1976, p. 216-217).

4.8. Conclusions: manipulation in von Wright's theory

Von Wright means to ground the counterfactual conditionals, that are implied by causal regularities but not by accidental ones, and can therefore be used to distinguish between these, in action counterfactuals—statements about things we *could* have done instead of what we *chose* to do on some occasion. Evidently, von Wright takes these action counterfactuals, stemming as they do from our beliefs about ourselves as free agents, to be immediately and pre-theoretically familiar to us, and therefore available as grounds in an analysis, even of causation and laws of nature. This perspective, which appears connected to the Wittgensteinian notion of internal relations (see e.g.: McCann 1978, p. 92; Stoutland 1982, p. 60; Tuomela 1982, p. 17), is unusual in the

philosophy of science, if not among causal manipulationists, specifically. The more common expectation has been that if nature is governed by laws at all, causal or otherwise, then these are objective, and in particular independent of the capacities or beliefs of humans or other creatures. On this naturalistic view, if there is an analysis, then it is agency that must be explained in terms of objective natural facts and laws, and not the other way around. Perhaps we could characterize the difference as between a default first person perspective, and a third person perspective, on agency. These differences in expectations have thus made von Wright's theory unpalatable or even unintelligible to many.

But taking the freedom of human action as a given can be theoretically powerful for the purposes of solving problems with causation. If the assumption that agents act freely implies that their actions do not have external, natural causes, then this satisfies the principal condition for the action being an *intervention*. This is the condition that the action is exogenous to the causal system that is being manipulated. That is to say, the action does not have as a cause any event that affects, or occurs in, the system. (This will be a rough-and-ready picture of what an intervention is—we shall develop it in detail in the chapter on James Woodward's interventionist theory of causation.) If the action has no cause at all, this condition is obviously satisfied. And if exogeneity holds, and there is a correlation between the action and a remote event, and we exclude the possibility that this correlation is a coincidence, then the action can neither have that remote event as its cause, nor can it have a cause in common with that event. Von Wright insists that actions do not have causes, so this argument is *prima facie* available to him. Moreover, if a p-event is the *result* of the action, and a q-event is the correlated remote event (what von Wright calls a consequence of the action), then since, on von Wright's theory, it follows logically from the p-event being the result of the action, that this event would not have occurred without the action, and that it therefore has no external sufficient cause, we can (given one last condition, that I will also postpone discussion of) conclude that the p-event is the cause of the q-event, just as von Wright claims that we can. What, finally, makes *any* event a cause of some effect is that it is an event of a type that can be manipulated (or that is composed of manipulable types of events) and if it *were* manipulated, the above conditions would hold. Or, rather, this *would* have been the case, had it not been for von Wright's compatibilism. Here I want to argue that von Wright's compatibilism, with respect to the relation between agency, and the existence of causal accounts of the bodily movements that are essential parts of our actions, is a fundamental problem for his manipulationist account of causation.

Von Wright's compatibilism amounts to the proposal that the freedom of our agency does not depend on our basic actions, *viewed as bodily movements*, lacking event causes. He says:

> In every action bodily behavior is also involved. This consists in the moving of limbs and other parts of the body, or in restraining such movements. The movement and not-movements are the causal consequences of muscular activity, i.e., of the

> contraction and relaxation of muscles in the body. Muscular activity is, in its turn, causally related to processes and states in the neural system of the agent. Neural activity may be caused by external and internal stimulation of the nervous system. The occurrence of the stimuli may also have causes, *ad infinitum*. (von Wright 1989, p. 810.)

But von Wright does not take this causal story to be an explanation of the action, since free agency is required to understand causation in the first place. That there *is* a causal story about those bodily movements that are actions when understood intentionally is then a contingent fact about them—it need not be so. This may well be a coherent view, as far as it goes, but it wreaks havoc on both the causal analysis and the logic of causal discovery. Because, wherever causes rank in the metaphysical or conceptual order of things, if intentional bodily movements have causes under *some* understanding of them, then we cannot exclude, as von Wright needs to do, the existence of a common cause that explains the co-occurrence of a p-event with a q-event, based on the fact that p was brought about by an action, nor that the q-event is not then in actuality the cause of the p-event (by being a cause of the action). This claim is grounded in a modern understanding of how our manipulations can confound our causal inferences, which will be stated explicitly in the chapters about interventionism, below. The theoretical force of assuming free agency—of the traditional sort—as basic, seems then limited to incompatibilist views of free agency.

CHAPTER 5

Price and Menzies: Causation as a Secondary Quality

5.1. Introduction

The landscape of the philosophy of causation changed radically between the publication of von Wright's manipulationist account in the first half of the 70's and the time when Huw Price and Peter Menzies proposed an agency-based understanding of causation as a secondary quality, in the early 90's. The philosophical mainstream in the intervening period had not been concerned with manipulationist or agency accounts of causation, but preoccupied mainly with theories stemming in one way or another from the regularist tradition.

J. L. Mackie's INUS analysis was published in 1965 and widely discussed thereafter (Mackie 1965). It was a regularity account, that defined a cause as an insufficient but necessary part of one sufficient condition (out of many possible such conditions) for the effect. Despite certain problems in the original presentation (see in particular Kim 1971), the INUS analysis was broadly influential on regularity treatments of causation, and Mackie developed his theory further in *The Cement of the Universe* (Mackie 1980). Mackie's approach was eclipsed by David Lewis's counterfactual theory of causation, first published in 1973 (Lewis 1987b), which would prove to be the dominant theory in analytic philosophy during the 80's and early 90's. This theory analyzed the causal relation in terms of the counterfactual conditionals the relation was taken to support, which in turn were understood according to Lewis's theory of these counterfactuals (Lewis 1973). Lewis originally introduced his theory as an alternative to current regularist theories, but because he has a regularist theory of laws of nature—the Best System theory (e.g., Lewis 1973, sect. 3.3)—and these laws are involved in determining the truth values of the counterfactuals that ground causal claims, it can be understood as belonging to the class of regularity theories. Many aspects of the logic of the causal relation were highlighted within the counterfactual framework, and these aspects were investigated in great detail over the coming decades, by Lewis and others. Among them were questions about the transitivity of the causal relation, and in particular cases of causal overdetermination and different kinds of "causal preemption" (see Paul and Hall 2013, ch. 3). Another approach was Wesley Salmon's mark transmission theory of causation (Salmon 1984). This theory focused on the propagation of causal influences in a process, rather than on a causal relation between events. A causal process is then one that can transmit a "mark" throughout its existence. A further development of this theory, that identifies the mark specifically

with a conserved quantity, has been defended by Phil Dowe (Dowe 1992). One question that this sort of theory brought particular attention to, is whether absences, omissions, and non-occurrences of events can be causes, as they tend to require special treatment in the causal process account (e.g., Schaffer 2004). A distinctly non-Humean way of understanding causation was promoted by Nancy Cartwright at least from the publication of her *How the Laws of Physics Lie* (1983). While Cartwright famously rejects any attempt to *define* causation, taking a pluralist stance on the question what causation is, she defends a notion of "causal powers" as intrinsic in objects. (The causal powers approach has recently received a more detailed treatment in Stephen Mumford and Rani Lill Anjum's *Getting Causes from Powers* (2011).) Probabilistic causation had received continued attention, too, in for example Skyrms 1980 and Eells 1991, and David Lewis had also contributed there (see the postscript to Lewis 1987d). Wesley Salmon's mark transmission theory, and D. H. Mellor's fact theory of causation (Mellor 1995) were also probabilistic treatments.

Thus, while most of these perspectives on causation have precursors in the philosophy of causation of the preceding generations, it is fair to say that the scope—both the width and the depth—of this branch of philosophy had grown significantly between the mid-70's and early 90's, and this is the backdrop against which Menzies and Price's updated agency-oriented proposal should be seen.

In a list of manipulation theories of causation, the paper "Causation as a Secondary Quality" by Peter Menzies and Huw Price is usually referenced (1993). The main focus of this paper is to suggest that a manipulationist theory of causation (they prefer the term "agency theory") can survive what they call the "stock objections" against such theories, if causation is viewed as a *secondary quality*. The strategy is to compare how a dispositional account of color fares against corresponding objections to it, and they claim that, as we don't think that such objections are fatal for the dispositional theory of color, nor should we regard it as an obstacle to the agency account of causation. The presentation of the theory itself is brief in this paper, and for further clues as to its details we will have to turn elsewhere. Menzies and Price predictably do not agree fully on every question about how causes or theories of causation should be understood. Huw Price had previously defended an evidential decision theory from certain arguments to the effect that causal decision theory is required, by employing what he called "agent probabilities" (Price 1991). This is one question where the views of Price and Menzies diverged, the latter holding that a causal decision theory is needed (Menzies and Price 1993, p. 190). Price has also recently said that "our philosophical dispositions were always a little way apart: Peter tended to be more of a realist, and more of a metaphysician, than I was (or am)" (Price 2017, p.74). Huw Price is moreover the one who has elaborated the most on the views expressed in "Causation as a Secondary Quality," both before and after its publication. So, in the interest of depth and focus, as we try to understand the theory beyond what is said in that article, I will concentrate on Price's views. But first I'll present a sampling of relevant ideas by Peter Menzies.

5.2. Peter Menzies on laws of nature and causation

Peter Menzies contributed a chapter to *Ontology, Causality and Mind: Essays in Honour of D. M. Armstrong* (Bacon et al. 1993). In it he proposed a theory of laws of nature and nomic necessity, in light of what he took to be intractable problems with both Lewis's and Armstrong's theories.

> I shall present a new theory of laws that does not force physical modality into the Humean straightjacket nor postulate irreducible necessary connections. The theory will appeal to a primitive concept of modality which we must all possess in virtue of being decision-making agents. This modal concept is that of the possible courses of events within an agent's control or, in other words, the possible outcomes that an agent could bring about by performing an action. (Menzies 1993, p. 195-196.)

And later.

> Even if I choose to open the window and so bring about the first course of events just mentioned, the second course of events was within my control at the time of decision. It is this sense of there being many alternative possible courses of events within one's control at the time of deliberation that lies, I shall argue, behind the modal character of laws of nature. (Menzies 1993, p. 208.)

If you have read the previous chapter, this theory of laws will sound eerily familiar. It is essentially the same suggestion that von Wright offers to ground the distinction between accidental and nomic regularities, that is then employed in his causal analysis. The similarities are substantial beyond the quoted passages. The notion of "possible courses of events," constructed out of situations that are described by atomic propositions, seems to correspond closely to the branches in von Wright's systems. Menzies, too, connects the theory to a modal logic, that uses special operators for experimental possibility and experimental necessity, and he thinks that "both forward-tracking counterfactuals and laws of nature must be explained in terms of the logically prior concept of an experimentally possible course of events" (Menzies 1993, p. 217). But there are also important differences, in the exposition at least. Since Menzies does not use an unnecessarily tensed formal language, does not use an idiosyncratic interpretation of "physical possibility" (such that there are physically impossible causal possibilities), and explains his proposal in the context of the theories by Lewis and Armstrong, it is by far the more accessible presentation. (That this part of von Wright's theory was not very accessible is perhaps indicated by the fact that Armstrong in his response calls Menzies's suggestion "an interesting addition to the currently available accounts of laws" (Armstrong 1993, p. 231).) I think we might want to call this theory the von Wright-Menzies theory of laws of nature.

In "The Role of Counterfactual Dependence in Causal Judgements," Menzies adds a distinction between "default" and "deviant" counterfactuals to a

modified version of David Lewis's counterfactual analysis of causation (Menzies 2011). (An earlier use of this distinction can be found in for example Hall 2007.) This distinction, in turn, is connected to judgments about what are the normal and abnormal developments of events, in ordinary human psychology. Employing an informal notion of "intervention," as a generalization of "the most important intuitive feature of an intentional action, which is that it represents an independent, exogenous causal influence on the course of events," Menzies now says that "a counterfactual is a *deviant counterfactual* if its closest antecedent-worlds are ones in which the antecedent is realized by an exogenous intervention; and [...] a counterfactual is a *default counterfactual* if its closest antecedent-worlds are ones in which the antecedent is realized in the normal course of events in the absence of any intervention" (Menzies 2011, p. 199). Finally, he proposes truth conditions for causal claims:

> A state of affairs c causes a wholly distinct state of affairs e if and only if (i) if c were to obtain, e would obtain; and (ii) if c were not to obtain, e would not obtain, where (i) is a deviant counterfactual and (ii) is a default counterfactual. (Menzies 2011, p. 200.)

Menzies relates his informal notion of intervention explicitly to the technical definition of "intervention" in the works by Judea Pearl (2009), Peter Spirtes, Clark Glymour, and Richard Scheines (Spirtes et al. 2000), and James Woodward (2003). Menzies's account as I have summarized it above may invoke a suspicion of circularity, but I think that Menzies didn't consider this to be an objection. Woodward had criticized the theory put forth in "Causation as a Secondary Quality" in his *Making Things Happen* (2003). In his review of that book, Menzies said that

> many of Woodward's criticisms [...] strike home. I am persuaded that it is much better to formulate a manipulability account that eschews reduction by using a notion of intervention that bears its causal character on its face; and that [it] is better to avoid anthropocentrism by appealing to a notion of intervention that is general enough to apply even in cases where human manipulation is physically impossible. (Menzies 2006).

He nevertheless had some doubts as to whether Woodward succeeds in delivering on his realist ambitions, that we shall return to when we discuss Woodward's theory in chapter 7.

In the rest of this chapter, I will mainly rely on the "Pricean" understanding of the theory offered in "Causation as a Secondary Quality."

5.3. Agent probabilities

While the focus of "Causation as a Secondary Quality" is the idea that causation under an agency theory should be understood as a secondary quality, the core of the theory itself is the concept of *agent probability*. The authors emphasize that the central difference between their theory and the earlier ones

by Collingwood, Gasking, and von Wright is this probabilistic treatment of causal dependence (Menzies and Price 1993, p. 189). The reason for this choice is to accommodate indeterministic causation. Menzies and Price note that deterministic dependencies can be treated as a limiting case in the probabilistic treatment.

The traditional, non-agency oriented probabilistic theory of causation, that harks back to the works of Reichenbach and Suppes, was an analysis of general causal relations in terms of statistical dependencies. The starting point was to take A to be a cause of B only if (i) $Pr(\mathtt{B}|\mathtt{A}) > Pr(\mathtt{B})$. But this dependence is symmetric, such that if $Pr(\mathtt{B}|\mathtt{A}) > Pr(\mathtt{B})$ then $Pr(\mathtt{A}|\mathtt{B}) > Pr(\mathtt{A})$, too. Thus, to distinguish cause from effect, Reichenbach added the condition that A is a cause of B only if (ii) A precedes B in time. Even given this addition, it can be the case that two types of events A and B are correlated, and A precedes B, even though A is not a cause of B. This happens when both A and B are effects of a third event (type) C. To exclude such cases, Reichenbach added yet another condition: there is no event C, occurring earlier than A, and such that conditional on *it*, the probabilistic dependence between A and B disappears. I.e., A is a cause of B only if (iii) there is no C earlier than A and such that $Pr(\mathtt{B}|\mathtt{A}.\mathtt{C}) = Pr(\mathtt{B}|\mathtt{C})$. In such a situation, C is said to "screen off" B from A (see the introductory chapter). Suppes's treatment was essentially similar to Reichenbach's. This solution then relied on the time ordering condition of which many philosophers have been skeptical. Moreover, Nancy Cartwright had argued in *How the Laws of Physics Lie* (1983, ch. 1) that getting the description of events right in condition (iii) depends on causal notions, and that the theory could therefore not be considered a successful reductive analysis. Cartwright concluded that causation was not reducible to statistical dependencies, but also that it cannot be eliminated from our ontology, since it manifests the vital difference between effective and ineffective strategies. In "Agency and Probabilistic Causality," Huw Price argues in response to these claims, that if we take agency and effective strategy as basic notions in the theory, we can have a successful probabilistic theory of causation (Price 1991). The way agency is introduced into the theory is in the form of *agent probabilities*. Price associates this approach with a "projectivist" stance on causation, and a view of causation as a secondary quality (Price 1991, p. 172), a perspective pursued further in "Causation as a Secondary Quality."

An agent probability is the probability of some event B conditional on the occurrence of an event A, where A is a *free act* by an agent. Menzies and Price express this as $P_{\mathtt{A}}(\mathtt{B})$. They then take A to be "a means for achieving B" if and only if $P_{\mathtt{A}}(\mathtt{B}) > P_{-\mathtt{A}}(\mathtt{B})$ (where "$-$A" denotes the non-occurrence of A). This defines the means-end relation that underlies the concept of causation in the theory. Now, given that the symmetry of probabilistic dependence mentioned above is a logical consequence of the axioms of probability theory alone, it must be expected to be present also when the events in question involve free acts by agents, so we want to know more about how Menzies and Price intend to resolve the asymmetry problem here. Specifically, how do we know that B is an *effect* and not a *cause* of A. They intend to make this distinction by

making the question of a means-end relation a *decision theoretical* issue, which then naturally involves free acts as a special class of events. The core idea concerning agent probabilities is that an act raises the probability of an effect of that act, and thus gives the agent reason to perform the act if she desires to realize the effect, but the act does *not* raise the agent probability for any *cause* of the act—since this would conflict with the presumption that the act is a matter of free choice. Therefore, there is no such probabilistic dependence that could give the agent a reason to perform the act when realization of its *cause* is an overriding preference. This leads immediately into the Newcomb problem in decision theory, and the bulk of "Agency and Probabilistic Causality" indeed concerns a defense by Price of evidential decision theory in the context of this problem. (Newcomb's Problem was introduced in Nozick 1969. For an overview, see Weirich 2016.) There is, then, a fundamental difference between statistical generalizations of merely observed behaviors of others (or of other events in general) and statistical generalizations involved in decisions about how to act. The following example illustrates and intuitively motivates the difference. Observing that someone enters a doctor's office gives us good grounds for raising our subjective probability for that they are ill—it is just a statistical fact (let's suppose) that people who enter doctor's offices are on average less healthy than those who don't. This would then be a matter of a regular Bayesian conditioning on the observed event. But that doesn't mean that deciding not to enter a doctor's office gives us grounds for thinking that we are now healthy. The very same statistical fact employed in the first inference is irrelevant as grounds for deciding how to act in the second case. The important difference between the case of observed behaviors of others and the case of making a choice as to one's own actions lies in the fact that in the first case an observed action can be an *effect* that constitutes evidence for its *cause*. Conditioning instead on a presumed free act in the case of decision making is then taken to block such "causal backtracking" inferences. This is taken as well to resolve the issue of excluding spurious correlations between `A` and `B`, attributable to a common cause. The general principle is familiar from von Wright's theory, although it is more explicitly stated by Menzies and Price. Taking `A` to be a free act amounts to imagining "a new history for [`A`], a history that would ultimately originate in one's own decision, freely made" (Menzies and Price 1993, p. 191). They also say that `A` is supposed to be realized *"ab initio*, as a free act of the agent" (Menzies and Price 1993, p. 190). Thus, since `A` is assumed to be a matter of free choice and thereby to have no preceding cause, a probabilistic dependence between `A` and `B` cannot be due to a common cause of them. Price expresses this point also in "Agency and Probabilistic Causality":

> To introduce the agent is in effect to assume an independent causal history to the event `A`. Those probabilistic correlations that survive this assumption seem to have claim to be counted as genuine effects of `A`. (Price 1991, p. 169.)

Menzies and Price then flesh out their proposal in relation to what they call four "stock objections" to manipulationist theories of causation.

5.4. Causation as a Secondary Quality

The four objections that Menzies and Price respond to are familiar to us. I quote their descriptions in full (Menzies and Price 1993, p. 188):

> 1. *Agency accounts confuse the epistemology of causation with its metaphysics.* It is widely conceded that experimentation is an invaluable source of evidence for causal claims; the objection is that it is a confusion to suppose that the notion of agency should thereby enter into the analysis of causal claims.
>
> 2. *Agency accounts are vitiated by circularity.* It is argued that the *bringing about* is itself a causal notion, and that this introduces a vicious circularity into an agency account
>
> 3. *An agency account cannot make sense of causal relations between events which are outside the control of any agent.* For example, it is argued that such an account cannot make sense of the claim that the earth's revolution around the sun causes us to experience the seasons
>
> 4. *Agency accounts make causation an unacceptably anthropocentric phenomenon.* Agency accounts are said to imply what is obviously false, namely that there would be no causal relations if there were no human agents (or different causal relations if there were different human agents.)

In the preceding chapters we have examined these issues, and noted in particular that the circularity objection (2) in fact may not strictly speaking fit any of the earlier theories. Although only von Wright addresses this point explicitly, it would seem as though both Collingwood and Gasking, too, do avoid circularity of this sort, whether it is by design or not. It has nevertheless been a common objection. As to the question of unmanipulable causes (objection 3), all of the earlier theories address this issue out of the gate in one way or another, and it has in fact not been a strong objection to the theories, as far as I have seen. Finally, the anthropocentrism charge (objection 4—we have called it "anthropomorphism" elsewhere) seems apt from a causal realist and objectivist perspective. Nevertheless, given that the goal of the earliest theories can be seen as showing how "cause" can be taken as a meaningful notion, while we at the same time accept that (as argued by Russell) causation is not an objective feature of the world, by instead connecting causation to particular human interests and perspectives, it seems problematic to view the anthropocentrism as a *failure* of the theories. The disagreement would then seem to be located at a different level, i.e. the question whether causation indeed is an objective feature of the world or not. Let us now turn to Menzies and Price's responses to these challenges.

These objections have, according to Menzies and Price, "seemed to show that agency cannot, in principle, play a constitutive role in an account of causation" (Menzies and Price 1993, p. 192). Their goal is then, it would seem, to argue to the contrary, that agency *can* play a constitutive role in a theory of what causation is. The strategy is to suggest that the agency account of causation may fruitfully be equated with viewing causation as a secondary quality. A dispositional theory of color is used as an example of a largely uncontroversial case where the presence and nature of a property depends on features of the subjects experiencing it. "Few philosophers would dispute that an adequate account of colour will need to make some reference to human perceptual states or capacities—that colour is a secondary quality, to use the familiar terminology." (Menzies and Price 1993, p. 188). The dispositional theory of color that Menzies and Price use can be summarized as, "an object is red, say, just in case it would look red to a normal observer under standard conditions" (Menzies and Price 1993, p. 192). They now propose that, to the extent that the dispositional color theorist has acceptable responses to these four objections, so, too, does the agency theorist about causation.

5.4.1. Metaphysics and epistemology confused. It is clear that, if we can discover causal relations at all, manipulation, and in particular experimentation, is the primary way by which we may do this. The objection now is that the manipulationist confuses this epistemic fact with a fact about what causation *is*, or about the content of the cause concept. Menzies and Price call this the "verificationist fallacy": "after all, it is the cardinal sin of verificationism to suppose that the means by which a statement is verified or tested determines the meaning of the statement" (Menzies and Price 1993, p. 192). Their response to this objection is that theirs is not a verificationist theory: they do not relate the meaning of causal claims to conditions for their verification or warranted assertion. What they claim is that causation is a secondary quality, and they compare this to the dispositional theory of color. This theory makes the concept of red external to the objects having that property, and related in its constitution to a particular human response, that of looking red. This has epistemic implications, since the best evidence for something being red is that it looks red, but that evidence is not incontrovertible, since something that isn't red can still look red to someone under non-standard conditions. The same is true of causation under the agency theory. Relative frequencies observed under manipulation are the best evidence for the presence of a causal relation, but this observed frequency can deviate "by chance from the true mark" (Menzies and Price 1993, p. 193). This makes any empirical evidence for a causal relation defeasible in principle and—as I understand Menzies and Price—sufficiently decouples the metaphysics from the epistemology in their theory.

5.4.2. Circularity. We have discussed the claim that manipulation theories are circular in how they define or understand causation at length in the previous chapters. To repeat, the objection points to the commonly accepted view that to bring an event or state of affairs about is to *cause* it to occur or be the case, and thus "bring about," or any of its synonyms, cannot be used in a

definition or explanation of "cause." In a sketch of the manipulationist theory, we may say that `A` is a cause of `B` just in case bringing about `A` is a means for an agent of bringing about `B`. In Rosenberg's critique of Gasking's theory (see sect. 3.4) , he described two ways in which manipulation theories can be accused of circularity along these lines, reflected in the two uses of "bring about" in the sketch.

First, if the theory requires that a human action *causes* `A` to occur, then this relation appears to remain unanalyzed in the account. But we noted that none of the earlier theories in fact suffered from this problem, when considered in full. Collingwood, who identified a cause$_2$ `A` of some `B` with a means by which we could bring about `B`, also clearly stated that the action involved is *not* our means for bringing about `A`. Consequently, the action is not the cause of `A`, according to the theory. In Gasking's theory, the logical properties of Gasking's conditions on the causal relation, that deliver the desired asymmetry of the relation, also exclude the possibility that a "manipulative technique" is the cause of its immediate result (see sect. 3.2). Finally, von Wright makes clear that he does not consider an action to cause its immediate result, since he takes the action and this result to be logically connected, and therefore not independent in the way required of cause and effect. (We also briefly noted that this conclusion doesn't necessarily follow, in 4.2.)

The second way circularity can enter the theory is in the claim in the definition that the cause brings about its effect, which seems to mean that it causes the effect. This was true of Gasking's formulation, but it is easily fixed by reformulating the condition in terms of the effect occurring or being realized when the cause is. Von Wright states his condition in terms of the sufficiency of the cause for its effect in the history of possibilities, so does not expose himself to this criticism.

Menzies and Price's novel response to the circularity charge appears to circumvent both of these possible objections, and moreover offers clues as to how causation and agency are conceptually connected on their view. They note that the same sort of circularity can be seen in the dispositional theory of color. If being red is defined as looking red to a normal observer in standard conditions, then clearly the concept "red" occurs on both sides of the definition. But Menzies and Price think that there is no vicious circularity here, because our understanding of "looking red" is acquired directly through experience. That is, "looking red" has an ostensive definition.

> [A] novice can be introduced to the concept 'looks red' by being shown samples of red: the salience of the redness in the samples and the novice's innate quality space should suffice for him to grasp the fact that the samples look alike in a certain respect. (Menzies and Price 1993, p. 194.)

They suggest that the situation is comparable in the case of the agential notion of bringing something about, in the sense that our understanding of the notion of accomplishing one thing by means of doing something else is acquired through personal experience, from an early age.

> We might say that the notion of causation thus arises, not as Hume has it, from our experience of mere *succession*; but rather from our experience of *success*: success in the ordinary business of achieving our ends by acting in one way rather than another. (Menzies and Price 1993, p. 194.)

Thus, if "bringing about" is understood ostensively through experience, then it would seem that its use anywhere on the right hand side in the analysis of causation is accounted for. This also provides a straightforward illustration of how we might take the suggestion that there is a conceptual connection between agency and causation. We acquire the agential notion of "bringing about" directly through experience, and then, as indicated by the definition, extend it to situations that do not involve our actual agency. The next step is to understand how we extend the cause concept further still, to situations that couldn't be manipulated.

5.4.3. The problem of unmanipulable causes. As I mentioned above, the question of how we can extend the causal relation to events that could not be manipulated is in no way an outstanding issue in earlier manipulation theories. Collingwood argues explicitly that there simply are no such causes. Gasking relates unmanipulable causes to manipulable ones based on the types of events they are, in a way that at least in many cases seems to allow for a classification in non-causal terms. (E.g., some events that are instances of water changing state from solid to liquid are caused by events that are instances of increase in regional temperature. Some of these events are of a scale or occur in such environments that we could not artificially realize the cause of the melting, but in some other cases we can.) Von Wright furthermore allows that a cause may be *composed* of events of a manipulable type. Nevertheless, by relating this issue to a corresponding problem with the dispositional theory of color, Menzies and Price gives it a new spin.

There are things, that are not observed, and that we want to say have a certain color, although the counterfactual "if that object had been observed by a normal observer under standard conditions, then it would have looked that color to the observer" either cannot be straightforwardly evaluated or comes out false. It might be, for example, that an object has a certain color *in virtue* of being in an environment such that no normal observer could exist there. So, if we hop over to a physically possible world in which a normal observer is watching this object, the environment of the object in that world will necessarily be different in such a way that the object doesn't have the color we want to ascribe to it in *this* world, and the dispositional condition therefore fails. An example of this is some part of the inside of the sun, which we would want to say, based on its physical characteristics, emits light that would look a certain color to someone if it were emitted in an ordinary environment. But the relevant physical characteristics depend on the fact that this is a part of the inside of the sun, that is exposed to all of the gravitational effects there—it can't be separated from them and retain this color, and no normal observer can exist in this environment. Thus, for an understanding of "that object" that is rigid enough to preserve its supposed color, the antecedent of the counterfactual

conditional is not realized in any physically possible world. To sum up, it is sometimes physically impossible for a normal observer to observe some object that we nevertheless want to say has a certain color. Analogously, there are cases of causation such that it is not physically possible that the cause be brought about or prevented by a manipulation.

The foregoing example invites the suggestion that we look for our observer among the physically impossible worlds. This would then amount to introducing an "ideal observer," that is not limited by what conditions allow for human presence and observation. Likewise in the theory of causation, we could introduce an "ideal agent," not subject to human limitations on manipulation. Menzies and Price, however, think that this move will not suffice, because there are other examples that they take to show that a straightforward counterfactual criterion cannot ground either a dispositional theory of color or of causation. These are examples of so-called *finkish dispositions*. A finkish disposition is one "which vanishes when it is put to the test." The example in Menzies and Price 1993 (p. 196) is the chemical substance rhodopsin. The chemical surface properties of rhodopsin are those of something that reflects light that looks yellow under normal circumstances. But rhodopsin is moreover such that when it is exposed to light, its chemical properties change so that it instead reflects red-looking light. Menzies and Price claim that "it is plausible to say that rhodopsin has the surface colour yellow, even though it does not look yellow to a normal observer under standard conditions" (Menzies and Price 1993, p. 196). After providing a third problematic example—that of an object having a dispositional property that is masked by other properties, they suggest a unified solution to all three sorts of cases. The solution has in a way been hinted at in the descriptions of these problems. Our judgment that rhodopsin is yellow (if that is our judgment) is derived from our beliefs about the chemical surface properties of the substance. The same is true when we attribute a color to some part of the inside of the sun. Menzies and Price accordingly suggest a weakening of the dispositional theory of color to make use of this fact: an object is red if and only if it would look red to a normal observer under standard conditions, or if "it possesses intrinsic properties which are identical with or closely similar to those of an object" which would look red to a normal observer under standard conditions (Menzies and Price 1993, p. 197). The problem of unmanipulable causes should be resolved in the corresponding way. That is to say, like Gasking, Menzies and Price allow an event to be a cause if it belongs to a type of events that has some manipulable instances, and they furthermore want to make this classification based on properties of the events that are intrinsic and "essentially non-causal though not necessarily physical" (Menzies and Price 1993, p. 197). This brings us to their definition of a causal relation:

> **MP:** A pair of events are causally related *if and only if* the situation involving them possesses intrinsic features that either *(i)* support a means-end relation (as defined in terms of agent probability in the previous section) between the events, or *(ii)* are identical or closely

> similar to the intrinsic features of a situation involving "an analogous pair of means-end related events" (Menzies and Price 1993, p. 197).

The theory thus implies that when we are faced with a cause which could not be manipulated, we have inferred that it nevertheless is a cause based on the intrinsic, essentially non-causal properties that it (along with relevant parts of the background circumstances) shares with some cause which *could* be manipulated. One might wonder why Menzies and Price add that these intrinsic properties are "not necessarily physical." Perhaps a clue is provided by their example situation. We take the earthquake in San Francisco in 1989 to have been caused by friction between continental plates, although this friction is not something we can artificially bring about. Menzies and Price now suggest that the "paradigm example" of a situation that "models" the earthquake and surrounding circumstances, and that supports a means-end relation between the corresponding pairs of events, "would be that created by seismologists in the artificial simulations of the movements of the continental plates" (Menzies and Price 1993, p. 197-198). Here, the relevant intrinsic features shared between the earthquake and its model—I will assume that it is a computer model—may not be physical properties in any straightforward sense.

5.4.4. Anthropocentricity. Menzies and Price state the objection that the agency theory is too anthropocentric in terms of the relativity of the causal relation to the capacities of agents. In its most "naive" form, the objection is that where there are no agents, there can be no causation. This clearly does not follow from the theory, given the way the causal relation is extended beyond the class of manipulable events, described in the previous section. The more serious objection is then that the causal relation in any possible world depends on the contingent powers and capacities of the agents in that world, and in a possible world without agents, there is no causation. Menzies and Price first bring up the standard response to the corresponding problem for the dispositional account of color. There, the problem is that what is red in some world is relative to the perceptual capacities of normal observers in that world. The response, then, "is to rigidify the relevant dispositions, anchoring them to the perceptual capacities of the normal observers of the *actual* world" (Menzies and Price 1993, p. 199). Thus, some object is red in *any* possible world just in case it would look red to a normal *actual* observer, rather than to a normal observer in that object's own possible world. However, this move doesn't in fact eliminate the observer-relativity of the concept "red." Menzies and Price refer to David Lewis for an explanation. In a discussion about the dispositional theory of values, Lewis had said that "[t]he trick of rigidifying seems more to hinder the expression of our worry than to make it go away" (Lewis 1989, p. 132). Of course, what some merely possible world represents is one way things *could* have been. If things had gone differently than they did, in a certain way, then things would have been that way in the actual world. Thus, if observers in that case had possessed different perceptual capacities, then the very schema for fixing the extension of "red" that the dispositional theory proposes would have produced a different extension than it in fact has (and to insist, in accordance with the rigidifying tactic, that the true extension

of "red" would nevertheless be fixed by an application of the schema in what would then be a merely possible world seems odd). This solution would then be no more effective in addressing the agent-relativity of causation. Now, Menzies and Price on the one hand accept that causation is agent relative, but on the other hand also want to, and think they can, confirm the intuition that causation "is significantly more 'objective' than the usual secondary qualities" (Menzies and Price 1993, p. 200).

Here is how I understand Menzies and Price argument. The way that they extend the causal relation to unmanipulable events, described in the previous section, implies that it doesn't matter for the question whether an earthquake is caused by friction between continental plates that we in fact lack the capacity to manipulate the continental plates. Specifically this means that had we *had* that capacity, the causal relation would have remained unchanged in this instance. Likewise, if we had been weaker than we are, and incapable of manipulating some things that we in fact can manipulate, then by the same principle we could have extended the causal relation from things we could then have manipulated to things that we then couldn't have manipulated, but as a matter of actual fact *can* manipulate. Again, this would supposedly not affect the extension of the causal relation. This principle of extension, then, suggests to Menzies and Price that "agents with different capacities will nevertheless envisage the same range of *possible* causal relations" (Menzies and Price 1993, p. 200). Specifically, this goes for merely possible agents with greater or smaller powers of manipulation. They suggest that the only possible world to which the anthropocentricity objection applies is one in which there are no agents at all, and they do accept that in such a world there is no causation. (This holds for a possible world in which there are beings that are not agents, but merely passive observers.)

This, then, amounts to a cause concept that is noticeably objective, for denoting a secondary quality. In "Causation, Intervention, and Agency: Woodward on Menzies and Price" Huw Price modifies his response to the anthropocentricity objection (Price 2017). More on this later development in the next section.

5.5. Criticisms

In this section I will review several objections to "Causation as a Secondary Quality" made individually by Daniel Hausman and in particular James Woodward, and I will note Price's responses to Woodward's criticisms.

5.5.1. Getting the conditions right. James Woodward points out in *Making Things Happen*, that under the description of a free action that Menzies and Price provide, their conditions are not sufficient for establishing a causal relation (Woodward 2003, p. 126-127). Even if we assume that `A` is an event that is the result of a free action, in the sense of not having any causes beyond perhaps the agent herself or her intention to do `A`, it doesn't follow that if `A` correlates with `B`, then `A` is a cause of `B`. Woodward brings up the possibility that `A` correlates with another cause `C` of `B`, which then explains the correlation between `A` and `B`. Another possible source of error would be if the action that brings about `A` also brings about `B`, but not *by way* of bringing about `A`. This

possibility is then that the manipulation itself is a confounder. As an imagined example, consider the situation when you distribute a treatment in the form of a pill together with a glass of water to a group of patients, and compare the result to a group that gets no treatment. Maybe a significant increase in successful outcomes in the treatment group is not due to the substance that you are trying to test, but to some other substance in the pill, or the water. (This is why we want to try to expose the control group to exactly the same things as the treatment group, apart from the substance under test.) These possibilities, and the conditions that can account for them, are only treated explicitly in the interventionist theoretical framework that Woodward's theory employs, and that chapter 7 is devoted to.

5.5.2. Menzies and Price's solution to the unmanipulable causes problem. I think that Menzies and Price's solution to the problem of unmanipulable causes may easily evoke a sense that what fixes the causal relation isn't the means-end relation between events, but the intrinsic features of the events that Menzies and Price appeal to when they extend the relation to events that cannot be manipulated. Daniel Hausman and James Woodward have both made this observation.

In "Causation, Agency, and Intervention," Hausman suggests that agency in Menzies and Price's theory appears to play a part only in the formation of causal concepts and acquisition of causal knowledge.

> According to this defense [against the unmanipulable causes objection], `b` may causally depend on `a` even when realizing `A`'s is not an effective way, nor any way at all, of bringing about `B`'s. Is this revised account still an agency theory? The role of agency seems to be restricted to identifying the intrinsic conditions in virtue of which causal relations obtain. Whether causal relations obtain depends on whether those conditions are met. Agency matters only in the acquisition of causal notions and the discovery of causal relations. (Hausman 1997, p. S18.)

On a very similar note, Woodward presents a dilemma in *Making Things Happen*. He takes it that either Menzies and Price are proposing that the facts about the relevant intrinsic features, by which the causal relation is extended to unmanipulable events, are fully reducible to "facts having to do with our experience of agency and facts about noncausal relationships of similarity to situations in which manipulable causes and the experience of agency are present," but in that case they have provided no argument in support of this. Or, they mean rather to say that

> quite independently of our experience or perspective as agents, there is a certain kind of relationship with intrinsic features that we exploit or make use of when we bring about B by bringing about A. Moreover, because this relationship is intrinsic and can exist independently of anyone's experience of agency, it can also be present even when A is not in fact manipulable

> by humans. If so, I would claim that this is essentially the objectivist position regarding the connection between causality and agency that I have endorsed: considerations having to do with agency and manipulability help to explain why we developed a notion of causality having the features it does and play a heuristic role in helping to characterize the meaning of causal claims, and have considerable epistemic relevance when we come to test causal claims, but agency is not in any way "constitutive" of causality. (Woodward 2003, p. 125-126.)

Woodward adds, moreover, that it may not be possible to identify the relevant resemblances between a manipulable situation and one that isn't manipulable, that allows the extrapolation of a causal relation in the first case to the second, in a non-causal way. (This is a more sophisticated variant of one of Alexander Rosenberg's circularity objections to Gasking's theory—see sect. 3.4.) The problem is that 'small-scale models and simulations of naturally occurring phenomena that superficially resemble or mimic those phenomena may nonetheless fail to capture their causally relevant features because, for example, the models fail to "scale up"—because causal processes that are not represented in the model become quite important at the length scales that characterize the naturally occurring phenomena' (Woodward 2003, p. 125). Hence, the adequacy of models must in such cases be based directly on getting the causal dependencies right, it's not sufficient that they correctly exemplify some set of non-causal features of the situation and, consequently, Menzies and Price's approach does not eliminate the causal relations in favor of the means-end relation together with non-causal intrinsic features of situations.

Clearly, both Hausman and Woodward are arguing under the assumption that Menzies and Price mean to propose an agency theory of what causation *is*, and provide an analysis of causation in non-causal terms. This has been the common way of understanding these proposals, especially by critics—we will have reason to question it below.

In "Causation, Intervention, and Agency: Woodward on Menzies and Price," Huw Price responds to Woodward's objection (2017, p. 89-90). He there claims that Woodward's interventionist theory requires a principle of the same sort as the one Woodward criticizes for extending the causal relations to situations that we know we will never *intervene* in. This could then either be a principle that appeals to intrinsic non-causal similarities, which would concede that possibility to Menzies and Price, or it could be a principle based in some more general inference rules, perhaps appealing to such things as physical symmetries. Whatever is required to ground these, Price takes to be available in his own account. He thus concludes that Woodward is essentially in the same boat as him. But, as we shall see in chapter 7, the situations are not quite comparable. This is because, looking back at section 1.2, Woodward's theory deals with the unmanipulable causes problem by way of option 3 described there: he claims that there are no causes that are not intervenable. There is, in other words, a possible intervention on every cause, in some sense of "possible" appropriate for the theory. He thus does not employ any principle comparable to Menzies and

Prince's, by which the causal relation is extended to non-intervenable events. Certainly, we rely *epistemically* on some such inductive mode of inference for extrapolating experimental results to new situations, but this seems to me to be a different matter.

5.5.3. Realism and agent relativity. In his *Making Things Happen*—the main topic of our chapter 7—James Woodward presents his interventionist theory of causation. He calls his theory a "'manipulationist' conception of causal explanation" (Woodward 2003, p. 6)—but interventions are not there defined in terms of actions or agency. Interventions are defined in general causal terms, such that a successful manipulation can be understood as an example of an intervention, but perfectly natural events can also have these properties. In accordance with this, Woodward objects to the view in older manipulationist accounts that the truth values of causal claims depend on the beliefs, attitudes, or experiences of agents (Woodward 2003, p. 118). Woodward defends this objection by reference to how we use experiments. If the part of the causal relation that goes beyond a mere correlation is a "projection" of our beliefs and attitudes, then

> what sense can we make of experiments designed to distinguish the claim that X causes Y from the claim that they are correlated because of the operation of some common cause? Are such experiments simply roundabout ways of finding out about the experimenter's (or the scientific community's) projective activities? (Woodward 2003, p. 119.)

Rather, Woodward claims, it is a *presupposition* in deliberation about the outcomes of experiments that "if it is possible to change Y by intervening on X, then there must be an independently existing, invariant relationship between X and Y that the agent makes use of when she changes X and, in doing so, changes Y" (Woodward 2003, p. 119). Woodward thus rejects the idea that some part of the causal relation is determined by the agent's experience, beliefs, or attitudes, and thinks that it is true in particular that "a commitment to some version of realism about causation [...] seems to be built into any plausible version of a manipulability theory" (Woodward 2003, p. 120).

Huw Price responds to this objection in "Causation, Intervention, and Agency: Woodward on Menzies and Price" (2017). Price identifies the objection with the anthropocentrism complaint, described in the previous section. He first reiterates the claim from "Causation as a Secondary Quality" that this objection gets no traction when causation is viewed as a secondary quality, and emphasizes that the cause concept retains a great deal of objectivity in their treatment there. But he also puts forth a later development of his own view of the theory he and Peter Menzies offered in that paper, and I will focus on Price's later interpretation here.

Price now thinks that the cause concept is in fact more similar to the concept "red" than what was suggested in the original paper, in the sense that causation, too, is substantially relative to different kinds of agents and environments. He gives two examples of how the extension of the causal relation can be different to different agents. One is due to a possible difference in

external physical conditions. It depends on the theory that our perception of the direction of time is determined by the entropy gradient of our world—the future lies in the direction of increasing entropy. Agents who live in a (part of a) universe in which the entropy gradient is the reverse of what it is here, and who aligns the direction of the causal relation with the direction of time, as we do, will have a causal relation that is the reverse of ours. (Price discusses this possibility at length also in Price 2007, and more briefly in Price 1992b.) Price's argument in support of "causal perspectivalism" relies here on the asymmetry of causation ultimately being explained, not by an objective fact about the direction of the entropy gradient, but more deeply by the agents' asymmetry of deliberation—i.e., the asymmetry between what things in the world an agent must consider to be options for her to change, and what things she must consider fixed. Past things are fixtures, our available options lie in the future (at least normally). This would then be an example of a theoretically possible way in which causation is agent relative, in virtue of the asymmetry of causation being ultimately explained by the asymmetry of deliberation.

Another example of agent relativity is more "homely": when evaluating what are causal factors of some event, it is common to recognize that we will take some factors on which that event counterfactually depends as "serious possibilities" and others not, and count only the serious possibilities as genuine causal factors of the event. Woodward recognizes just this fact in *Making Things Happen*: "I believe that some relativization to 'serious possibilities' will be a feature of any plausible theory of causation" (2003, p. 56). An example commonly used to illustrate this concerns the death of a patient that was caused by the neglect of their attending doctor. Although that death may counterfactually depend to just the same degree on the corresponding inactivity of some other arbitrarily selected doctor at some other hospital, we wouldn't in our ordinary causal judgments include that inactivity among the causes of the death in question (Price 2017, p. 86).

Referring back to his "Causal Perspectivalism" (2007), Price describes deliberation about actions (for example in connection to an experiment) as essentially involving a division of the world into the mutually exclusive categories of *fixtures* and *options*. The options are the things that are "matters of deliberation," that is, the things which the agent takes herself to be able to bring about or not. The fixtures are "everything else—all matters of fact that are not held to be a matter of choice in the deliberation in question." He then observes that all of the known and knowable facts about the situation under deliberation must be part of the fixtures (quoting his own 2007 paper):

> it seems incoherent to treat something both as an input available to the deliberative process, at least in principle, and as something that can be decided by that process. Control trumps a claim to knowledge: I can't take myself to know that P, in circumstances in which I take myself to be able to decide whether P, in advance of that very decision. (Price 2017, p. 85.)

And (again quoting the earlier paper): "acting, or intervening, is a matter of fixing something not already fixed—of moving something from OPTIONS to FIXTURES [...]" (Price 2017, p. 86). Price then suggests that agents living in a world with a temporal arrow that is the reverse of ours would make different judgments about what things are fixtures and what things are options for them, and this would amount to different causal judgments. He also claims that in the case of judgments about what are serious and non-serious possibilities, this is again a matter of classifying some things—such as "the behaviour of distant strangers"—as fixtures, and others as options.

Consequently, Price now finds color and causation to be more comparable as secondary qualities. He denies that, as opposed to for color,

> for causation [...], there's only one possibility: one set of relations on which any creature capable of making the journey will inevitably converge. [...] I've argued that that's not the case. There are ineliminable contingencies in the causal case too—strikingly those of temporal perspective, though these are merely the most stark manifestation of something deeper, and elsewhere much more familiar. (Price 2017, p. 93).

5.5.4. Naturalism and philosophical anthropology. As we have already noted, in *Making Things Happen*, Woodward assumes that Menzies and Price intend "a noncircular 'analysis' of causation" in terms of the notions of "agency and manipulation by a human agent" (Woodward 2003, p. 123). This is an understandable assumption for several reasons: the "stock objections" that Menzies and Price's paper aims to dispose of are objections to the possibility of such a reductive analysis of causation in manipulationist terms; Menzies and Price describe these objections as having seemed to show that agency cannot "play a constitutive role in an account of causation" (Menzies and Price 1993, p. 192); and Price acknowledges that paper's "tendency to characterize the project of an agency theory in a metaphysical key" (Price 2017, p. 88). Woodward thinks that such a treatment "flies in the face of any plausible version of naturalism: it makes agency out to be a fundamental, irreducible feature of the world and not just one variety of causal transaction among others" (Woodward 2003, p. 123). While Price acknowledges in Price 2017 that "Causation as a Secondary Quality" is plausibly understood as being about the metaphysics of causation, he emphatically rejects this view in his more recent interpretation of the theory.

Already before the publication of "Causation as a Secondary Quality" Price called the view of causation he advocated "projectivist." He contrasted projectivism with causal realism, which aims to produce an account of causation as a "basic constituent" of the physical world, and compared this projectivism to Simon Blackburn's concept of projected predicates, in Blackburn's *Spreading the Word* (Price 1991, p. 160). Blackburn's theory is elaborate, but a crucial feature of it in our context is that the projection of "an attitude or habit or other commitment [...] onto the world," that occurs when we describe parts of the world in terms of projected predicates, does not constitute a *mistake* (Blackburn 1984, p. 170-171). If we call the latter view the Humean notion of

projection, then it is what for example Collingwood uses when he—very much in Hume's spirit despite his other misgivings—rejects a necessary connection between natural events, and explains the belief that there is such a connection as mere projection. Blackburn's theory of "quasi-realism" amounts to showing instead that there is nothing wrong with employing projected predicates in our descriptions of the world (Blackburn 1984, p. 171). Price then takes his account of causation to be much in this projectivist spirit. Price also suggests that "an attractive elaboration of a projectivist approach to causation" employs the fact that "the agent's perspective is something we all have—something that may thus be considered prior to the analytic task of understanding causation" (Price 1991, p. 172). By way of example, then, he compares this approach to that of secondary qualities. I think that we might broadly characterize these two ways of understanding projection as "Humean projection" and "Kantian projection"—the first taking projection to explain certain *false beliefs* about the world, the second taking it rather to explain how some *truths* about the empirical world are in fact agent (or more generally subject) relative. In "Causal Perspectivalism" Price expresses the nature of his investigation into causation with explicit reference to Kant:

> By identifying some key elements in those aspects of our epistemic and practical 'architecture' that seem essentially associated with causal thinking, I'll offer an abstract characterisation of what might be called the *causal viewpoint*: a distinctive mix of knowledge, ignorance and practical ability that a creature must apparently exemplify, if it is to be capable of employing causal concepts. My project is thus a kind of naturalized Kantianism about causation. It aims to understand causal notions by investigating the genealogy and preconditions of causal thinking; by asking what general architecture our ancestors must have come to instantiate, in order to view the world in causal terms. (Price 2007, p. 254-255.)

In particular the reference to "preconditions" of causal thinking may evoke thoughts of a Kantian transcendental argument for the necessary reality of these preconditions. In later accounts the focus seems to move even more towards the idea of providing a *genealogy* of the cause concept. This is connected to what Price calls "philosophical anthropology" (which he elaborates on also in Price 2010). He means

> explicitly to *disavow* that the project of the agency theory should be seen as *metaphysics* in the first place. Rather, it should be seen as what I have sometimes called philosophical anthropology: the task of explaining why creatures in our situation come to speak and think in certain ways—in this case, in ways that involve causal concepts. I think that this is one of a range of philosophically interesting cases in which the useful questions turn out to be questions about human thought and language, not questions about other aspects of the world (such as the nature of causation). (Price 2017, p. 75-76.)

Thus, Price agrees with Woodward that taking agency as a "metaphysical primitive" would be bad, but that this would constitute a misunderstanding of the proposal (Price 2017, p. 77). Importantly, Price also thinks that the "philosophical anthropology" approach to the agency account of causation is perfectly compatible with naturalism. Again, Price suggests that it is no more in conflict with naturalism than treating color as a secondary quality is.

> The general lesson is something like this. Many of our concepts are useful to us in virtue of contingent features of our own circumstances [...]. It is not surprising at all, from a naturalistic perspective, if some of our concepts reflect these 'located' features in essential ways—that is, roughly, in such a way that we cannot understand the concept in question except with reference to the feature in question. (Price 2017, p. 87-88.)

5.6. Conclusions: agency in Menzies and Price's theory

5.6.1. Global antirealism and primary qualities. There is an issue with Price's responses to the criticisms that I find makes evaluating his proposal exceedingly difficult. The criticisms miss their mark, it seems, because they mistake the nature of Menzies and Price's theory. It is not an investigation in analytic metaphysics—Menzies and Price are not, that is, concerned with giving an account of what causation *is* or is constituted of. Neither is it meant to be a "reductive *analysis* of the concept of causation" (Price 2017, p. 77). This is why Price thinks that the circularity objection in particular has no "bite." The source of my confusion lies in the fact that Price has proposed a *global* non-representationalist and quasi-realist theory of truth and meaning (see for example Price 2010). Up to a point, it might seem as though his agency theory of causation gives a treatment of the concept of causation that conforms to this general framework. He also thinks that it is the "genealogy" of concepts of things that is the "useful" question in general, rather than, say, the metaphysics of these things (Price 2017, 75-76). However, I think that this globally projectivist theory cannot really help us understand his agency theory of causation. The reason is that a globally projectivist theory cannot provide the distinction it seems that we should be interested in, between secondary qualities and *primary* qualities, since that theory interprets *all* quality concepts quasi-realistically. If, on the other hand, Price really holds that *every* quality is a secondary quality, then this suggests that there are arguments for the agent relativity of every such concept, mirroring those here given for causation, and consequently that the necessary condition that we use here to distinguish manipulation theories is present in the correct and useful theory of every quality concept. So, for example, something would have a rest mass only if that thing has a certain relation to manipulable things.

Certainly, every concept we have has a history that is intertwined with our interests and activities. For example, the concept of water is in this way historically related to our concepts of hydration and drowning, and the concept of weight is historically related to our concept of lifting. This may indeed be

important to an empirical genealogy of these concepts, but that Price is committed to the existence of primary as well as secondary qualities—and thus to a real distinction like this among actual qualities—is nevertheless also implied: firstly, the proposal that causation is a secondary quality seems rather less illuminating if everything is; secondly, Price describes his and Peter Menzies's theory as "proposing that the agency view should be regarded as taking causation [...] to lie on the 'secondary' side of the primary/secondary divide." So I will continue under the assumption that Price means that causation belongs to the class of secondary qualities, as opposed to the real and non-empty class of primary qualities. I take it that we can investigate this suggestion without involving any particular globally projectivist theory of truth and meaning that would apply to the interpretation of *all* claims about any qualities. (To involve that theory at this point now looks to me like a category mistake.) We then want to examine the arguments in support of the claim that causation, specifically, is a secondary quality. I will take these arguments as providing the support in this theory for the necessary condition, that we have identified as characteristic of manipulation theories of causation (see subsection 1.3). What is Price's arguments, then, for causation being a secondary and not a primary quality? I can see two such arguments. Neither of them seems to have an obvious equivalent for what we commonly have considered primary qualities (such as rest mass), further supporting our assumption that this is a distinction, in Price's theory, that can be made between real qualities that we find in the world.

5.6.2. Argument 1: Different agents can make different (true) causal judgments. The most recent argument is that differences in agents, their interests, or environments can result in different extensions of the causal relation for these agents. Price gave two examples, described in the previous section.

The first example concerned agents living in a world, or part of a world, with an entropy gradient that is the reverse of that in our (part of the) world. It is assumed that this gradient determines the direction of time for these agents, and that these agents line up the direction of causation in the same way. It's central to this argument in support of causal perspectivalism that the direction of causation isn't *defined* to be the direction of the entropy gradient, or the typical direction of the entropy gradient. Rather, Price argues that the better explanation for why causation has this direction is in terms of the asymmetry of agents' deliberations. The argument also hinges, it seems to me, on the claim that "however much we acknowledge that in [an entropy-reversed] universe our time-reversed cousins would see things differently, we can't imagine our own perspective shifting to align with theirs" (Price 2007, p. 277-278). That is to say, for there to be an agent relative concept of causation, our extension for this concept must differ from that of our time-reversed relatives. The details of Price's argument are subtle and complex, and well beyond the scope a very brief review like this. But if there is an immediate issue, from my perspective on these things, it's just this, that it's not clear how I would or should judge the situation in the thought experiment. When considering certain events in the

time-reversed world, do I think that a great number of porcelain shards had trajectories such that when they converged a coffee cup was brought about, which then shot up and landed on the kitchen counter—or do I rather think that the cup fell to the floor and shattered, but that time, and therefore also causation, is reversed in this world? I don't know. It's no help trying to figure out what my experience would be if I were to travel to this world, since I have no idea how my cognitive apparatus would interact with such an environment. Finally, it seems to me as though my hesitation on this issues may be a symptom of the fact that the argument relies on an agency view of causation in the first place. I am therefore at least not wholly convinced that this is an example of agent relative causation.

The second example referred to different judgments about what are "serious possibilities" with respect to factors on which the effects counterfactually depend. I think this argument for the agent—or rather subject— relativity of causation is more obviously successful than the previous one, but it is also more uncontroversial, as reflected in Woodward's embracing it, and the recognition in the philosophy of causation in general, of some pragmatic conditions for acceptable singular causal claims. It also doesn't seem to involve manipulations specifically, but just experiences. In the example, a particular event *counterfactually depends* on some other things, due to the physical laws, but these other things are nevertheless not considered to be serious possibilities, and are therefore not included among the causes of the event. I think the usual example introduces some potentially confusing elements in our judgment about it. That example was one where the death of a patient was caused by their doctor neglecting to administer some treatment. The death could also have been prevented by some other doctor, arbitrarily chosen from some other hospital, but we presumably don't judge that doctor's failure to act as a cause of the death. Clearly, this example involves a judgment about moral responsibility. We may reasonably think that the attending doctor's failure to administer the treatment made them responsible for the death, if it was their job to keep the patient alive. It is less likely that it was the job of the arbitrary doctor, so they have no responsibility for the death. Let's use instead another example, which doesn't involve a judgment about moral responsibility.

We may say that the lack of rain caused a certain season to fail to yield sufficient crops in a some area. We wouldn't say that the cause of crop failure was that oxygen and hydrogen in the atmosphere failed to spontaneously combine into water. This judgment appears to me not to be a judgment about the relevant causal laws governing the situation. We think that moist soil is required for a good harvest. There are several ways the soil can become moist, some more likely than others. We can differ in our judgments about the *probabilities* of these events, without differing in our judgments about general causal relations. Many singular causal claims, perhaps all such claims that are made outside of philosophical contexts, appear to occur as a response to some at least implicit question. These questions come with certain presuppositions. When we reject the claim that the cause of a certain crop failure was that water didn't spontaneously form out of the atmosphere, this can be understood as rejecting

the implicit suggestion that such an occurrence was an expected or hoped-for event on this occasion. (For a good discussion about this, see Menzies 2007.) But even if someone thought that spontaneous water formation was a "serious possibility," and therefore its absence a salient factor in explaining the crop failure, this doesn't seem to amount to a different judgment about what the relevant general causal relations are, that is, about the physics of the situation. Rather, this sort of difference in judgment about singular causal claims can be taken as reflecting a difference in judgment about the probabilities of events that the particular effect in question counterfactually depended on *in virtue of* the underlying causal laws, rather than a difference in judgment *about* those causal laws.

Note that, while it seems as though a pragmatic element is irreducibly involved in our judgments about singular causes in this way, this pragmatic element does not involve *manipulations* specifically. It's hard therefore to see how it defends a manipulationist account of causation rather than a more general subjectivist account. Also note that this sort of pragmatic condition on the acceptability of singular causal claims has been generally recognized across most theories of causation, even when it is not taken to condition what *exists* in the world, precisely because it is a subjective condition (see for example Paul and Hall 2013, p. 35). In other words, there has been an assumption in theories of causation that the subjective part of a singular causal judgment can be separated from its objective truth conditions. Rejecting this assumption is also what Collingwood does when he makes the truths of causal claims depend on the interests of individual agents.

5.6.3. Argument 2: The agency theory solves certain theoretical problems. For the reasons given in the previous section, it seems to me that the most straightforward argument for the manipulationist theory specifically—that is to say, for the necessary condition that we take to characterize this class of theories, and which says that every cause has a particular relation to something manipulable—also in Huw Price's presentation, is the by now familiar argument that only such an account can solve some commonly recognized theoretical problems associated with the causal relation. In particular, the account can provide a conceptual or theoretical distinction between cause and effect, and between causally and spuriously correlated events. This argument was presented by Price in "Agency and Probabilistic Causality" (1991), "Agency and Causal Asymmetry" (1992a), and "The Direction of Causation: Ramsey's Ultimate Contingency" (1992b), and it is outlined also in Menzies and Price's coauthored paper. The broad strokes of the argument are the same as in von Wright's theory, although it is made somewhat more explicitly in Price's and Menzies's and Price's accounts of agent probabilities.

To summarize, then, assume that an instance of the event type `A` is the result of a free act and thereby has a "history" that "ultimately originate[s] in one's own decision, freely made" (Menzies and Price 1993, p. 191). Price traces this assumption about the histories of free acts back to Frank Ramsey (Price 1992b, sect. 6). Then, `A` is a cause of `B` (under the prevailing background

conditions) if and only if A increases the probability of B. This rule is extended to events that can't be results of such free acts in the way detailed above.

Now, that we can draw this conclusion may seem to us to show that the "history" assumed to belong to A is its *causal* history. That A's history doesn't involve any causes external to the agent is what excludes both of the possibilities that A and B have a common cause and that B is the cause of A. Price may however not put it in this way. Rather, he might say that the rule is available to us even if we don't possess a concept of cause or of causal history. According to "Causation as a Secondary Quality," the rule is something we have direct experiential familiarity with. (The experience is that of "success" in actions.) On the other hand, Price seems to recognize that once we *do* have causal concepts, we would understand the validity of the rule by taking it to imply that A lacks external causes. Moreover, in his attribution of this idea to Ramsey, it is made explicit that the history is causal: "from the agent's point of view contemplated actions are always considered to be *sui generis*, uncaused by external factors" (Price 1992b, p. 261). Moreover, if we don't involve causal reasoning in our understanding of this rule, then it is hard for me to see why free acts are characterized in terms of their histories in this argument at all. When we experience success in actions, do we really involve the histories of the results of our free acts? Do we need *that* concept—the concept of an event that is *sui generis*—for acquiring, by ostension, our concept of a means to an end? Don't we just act, and certain things happen, that we may have intended to happen (success) or not (failure)? It seems to me that the assumption about this supposed history of the result of a free act is introduced in the argument precisely because it allows for the causal conclusion, since it implies that a free act lacks external causes. If so, then this argument for the agency account of causation itself *depends* on causal reasoning for its plausibility. But Price insists that circularity is not an issue in a theory of the type he here proposes, so we'll accept that for now.

I take it that the agency account presented in "Causation as a Secondary Quality" is taken to illuminate what it *means to say* that A is a cause of B. Therefore it seems to me as though it doesn't depend on whether the results of free acts in actuality *have* such histories as indicated by the "Ramseyan assumption" or not. The conclusion, that A is a cause of B, is something that follows under the given assumptions, whether these are true or false. This is an important difference to von Wright's presentation. Von Wright made both claims, in a way. He said that it was a category error to attribute causes to actions—but also that the bodily movements that are an essential part of an action does have natural causes. The latter compatibilist commitment blocks the wanted result in von Wright's theory, as we can then not conclude that the observed effect didn't have a common cause with that bodily movement, nor that it wasn't in fact the cause of the bodily movement. We can understand, I think, Menzies and Price's argument rather as articulating an *inference rule* that we actually have. If A is the result of a free act, thus conceived, and A makes B more probable, then we are entitled to conclude that A is a cause of B.

Thus, in supporting this inference, the concept of cause must have a certain content, which the inference rule brings to our attention. But note that there is no way to understand this situation that makes `A` being the result of a *free act* necessary for the conclusion. It's that it lacks a certain kind of causal history that is required. Imagine that `A` has no history at all. (It may for example be a fundamentally random occurrence.) Clearly, if `A` is such an event, and `A` makes `B` more probable, then we are entitled to the same conclusion: that `A` is a cause of `B`. (Or at least, any objection to that inference would be applicable also to the inference employed in the agency account.) So, even if the circularity present in characterizing causes in terms of free acts, and these in turn in terms of a certain kind of causal history, is not a problem because what is intended is not an analysis, we still appear to have no argument for the claim that all causes have a *necessary relation* of a certain kind to manipulable things. Recognizing this takes us one step towards James Woodward's interventionist theory, which does not appeal to manipulations, but just to events with the required causal properties. That theory seems to get the inference rule more precisely right—and we saw above that Peter Menzies in fact embraced this aspect of Woodward's theory, in favor of his and Price's original agency formulation.

5.6.4. To metaphysics or not to metaphysics. Recently, Price has responded to complaints that the agency account of causation that he defends is unacceptably anthropocentric and circular by insisting that it is not an analysis, and not a metaphysical account of what causation is. We saw above that the specific defenses against these criticisms given in "Causation as a Secondary Quality" were somewhat different to this, but Price had expressed the view already in 1992:

> [T]his is not an account of what causation *is*, but an account of how we come to speak in causal terms. (Price 1992a, p. 518.)

But we noted also that both James Woodward and Daniel Hausman did take "Causation as a Secondary Quality" to propose a theory of what causation is, and their criticisms—like those of every critic of manipulation theories that we have cited—were based on this assumption. How come? Is it only a matter of old philosophical habits refusing to die? This seems especially unlikely in the case of Woodward, who explicitly rejects these metaphysical traditions. I think rather that the presentation of the agency account invites an understanding of it as a theory that at the very least has substantial implications for what causation is. In the same 1992 paper, Price states that

> in a certain sense causal asymmetry is not in the world, but is rather a product of our own asymmetric perspective on the world. (Price 1992a, p. 513.)

As far as I can understand, the "certain sense" in question here can only be the *ordinary* sense, that has usually been employed in the metaphysics of causation: it is in Price's special, projectivist sense that causal asymmetry *is* in

the world—in virtue of being projected onto the world by us. Thus, this appears to be a metaphysical claim about what causation is "in the world." In general, any claim that some aspect of causation is a projection, *rather than* an objective feature of the world, is a metaphysical claim. Price's philosophical-anthropological theory about why "creatures in our situation come to speak [...] in ways that involve causal concepts" is therefore not compatible with any metaphysical theory that takes the asymmetry of the causal relation to be an objective (agent independent) property of it. Compare this to the way Spirtes, Glymour, and Scheines frame their treatment of causation:

> Views about the nature of causation divide very roughly into those that analyze causal influence as some sort of probabilistic relation, those that analyze causal influence as some sort of counterfactual relation (sometimes a counterfactual relation having to do with manipulations or interventions), and those that prefer not to talk of causation at all. [...] With suitable metaphysical gyrations the assumptions [made in this treatment] could be endorsed from any of these points of view, perhaps even the last. (Spirtes et al. 2000, p. 19.)

In his review of James Woodward's *Making Things Happen*, Clark Glymour appears to embrace a fourth option, namely that the causal relation (expressed in a certain way) is "an unanalyzed primitive relation" (Glymour 2004). It's unlikely that Glymour intends this as a contribution to the metaphysics of causation, but it shows that the interventionist framework he prefers is compatible also with at least some primitivist theories of causal powers (see Mumford 2009). Which theories of what causation *is* is Price's agency account compatible with? It seems to me that, the fewer these are, the more defensible it is to treat the agency account as a competitor on the metaphysical stage, just like most critics have.

5.6.5. Summary. I want to inject here that I have great sympathy for the criticism made by Menzies and Price in "Causation as a Secondary Quality" of the "passivist" empiricism according to which the only thing that could genuinely be involved in our discovery and understanding of how the world works is the particular order of individual events that march by in our visual field (Menzies and Price 1993, p. 202). I think that it's essential that we aren't just spectators—we are actors. Our question, however, is whether there is a convincing argument to be made in defense of the claim that *every* cause has a certain relation to something manipulable—the necessary condition I identified in the first chapter.

In this section I have argued that none of Price's arguments fully succeeds in this respect. I think that it's unclear whether the argument that contemplates agents in an environment with an entropy gradient that is the opposite to ours succeeds, because it's not clear to me whether this thought experiment implies a conceptual difference in the end. The argument relying instead on our admittedly differing judgments about what are serious possibilities, among the factors that an event counterfactually depends on, fails because, firstly, it doesn't involve manipulations in any obvious way, and secondly, because such

pragmatic conditions have regularly been combined with non-agency accounts of causation. Finally, I have tried to show that the argument that only the agency approach can account for the asymmetry of the causal relation and exclude cases of spurious correlation relies on a *causal* characterization of "free acts." This is not the circularity complaint; I mean to say that any event that fits the same causal description can do the job in the theory, and thus, causes do not need to have a special relation to free acts specifically.

Finally, I have questioned the effectiveness of Price's general defense against certain objections, where he denies that his theory is a metaphysical theory of causation. It seems to me that Price's proposal is metaphysically significant, as indicated by the fact that it is incompatible with every theory implying that, for example, the asymmetry of the causal relation is an objective feature of the world. It seems questionable, for this reason, whether there is anything wrong with treating it as a metaphysical theory. The suggestion here is, perhaps, that one cannot avoid doing metaphysics by fiat. And that the proper sign of a metaphysically innocent theory is that it doesn't conflict with at least a reasonable number of metaphysical theories.

CHAPTER 6

Interventionism in context: Causal Models

6.1. Introduction

This chapter provides some important context to James Woodward's interventionist theory of causation, reviewed in the next chapter. "Interventionism" has become the label of a kind of theory with two distinct heritages. "Intervention" is in these theories a technical term denoting an event with certain causal properties. This class of events is interesting in particular because manipulations in successful experiments are assumed to belong to it. By extension, the conditions on interventions also illuminate the meaning of a "natural experiment"—a naturally occurring situation that has some of the epistemically important properties of a true experiment. "Intervention" can therefore be said to be a generalization of "manipulation." Moreover, the theoretical role of interventions has much in common with the theoretical role of manipulations in the philosophical accounts of causation that we have discussed in the previous chapters. From this point of view, interventionism belongs to the history of these manipulationist theories in the philosophy of causation. We might call this interventionism's philosophical heritage. But an at least equally important influence on interventionism—some I'm sure would argue that it is much more important—is the history of certain statistical methods for formulating and testing causal hypotheses based on statistical data, that were developed beginning in the early 20th century.

Of particular importance are those methods that employ *causal diagrams*, usually called Structural Equation Modeling (SEM) or Structural Causal Modeling (SCM). Structural Causal Modeling is most closely associated with the work of Judea Pearl, and may differ from classical SEM mainly by being more explicitly causal in its interpretation, and by having a preference for Bayesian, non-parametric models. I will not pay close attention to these differences in this chapter, but call the general framework characterized by causal equations and causal diagrams, "causal models" or "SEM." Another important methodological framework, most prominent in econometrics, is that of *Potential Outcomes*. These traditions have mainly been concerned with identifying the conditions, or assumptions, under which we can infer causal information from purely observational data, while the efficacy of true experiments and manipulation for revealing causal relations has been taken as a given, and used as a model.

What follows here is a very brief overview of the scientific branch of the history of philosophical interventionist theories of causation, and the scientific

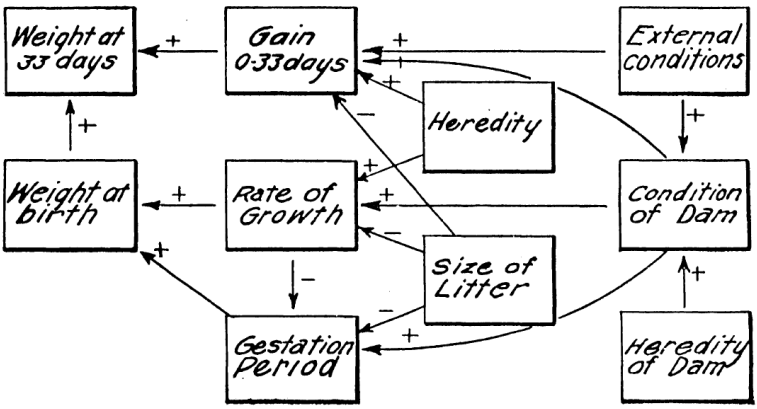

FIG. 1.—Diagram illustrating the interrelations among the factors which determine the weight of guinea pigs at birth and at weaning (33 days).

FIGURE 6.2.1. An early causal diagram from Wright 1921 (p. 560).

use of causal models, before we proceed in the next section to review James Woodward's theory.

6.2. A brief history of causal modeling

Historically speaking, causal modeling encompasses both *factor analysis* and *path analysis*. Factor analysis aims to identify some unobserved factors in a system under study, that explain the correlations of two or more observed phenomena, represented by variables in statistical data. Such unobserved factors are in turn represented in a theory by "latent" variables. This method traces back to Charles Spearman's psychological experiments, reported in Spearman 1904. There, Spearman concluded that the correlations among observed relative proficiencies in a number of intellectual activities, such as "in-school Cleverness, out-of-school Common Sense, Sensory Discrimination, and Musical Talent," could be explained by their dependence on a single factor: "All examination, therefore, in the different sensory, school, or other specific intellectual faculties, may be regarded as so many independently obtained estimates of the one great common Intellective Function" (Spearman 1904, p. 272).

Path analysis, in turn, was first developed by the geneticist Sewall Wright. The paths in question are paths of causal influence between observed quantities. Wright proposed "a method of measuring the direct influence along each separate path in [...] a system and thus of finding the degree to which variation of a given effect is determined by each particular cause" (Wright 1921, p. 557). Of special importance to later developments in causal modeling was Wright's use of directed graphs to represent the structure of causal influences between different factors in a system under study.

The structure in figure 6.2.1 is qualitative, showing only the paths of positive or negative causal influence between factors in the system, but not the

individual strengths, or "weights," of these influences. (Modern causal diagrams usually don't show whether influences are positive or negative either.) Wright presents a method for estimating path coefficients, attaching to each arrow in the graph. The calculation is based on information about correlations, and possible given certain assumptions, such as that we can hold some factors fixed while others are allowed to vary. The reliability of the estimation of path coefficients also depends on the causal hypothesis depicted in the graph being correct.

In general, then, a complete causal theory is expressed by a causal diagram together with a set of equations, one for each variable in the diagram, stating the quantitative (sometimes probabilistic) relations of causal influence between the connected variables. These equations have a special causal interpretation, and we will indicate that by calling them *causal equations*. The equations are also called "structural equations," and that a causal theory consists of a structure of quantitative dependencies, expressed by a set of such equations, gives this method also the name "structural equation modeling."

An important part of causal modeling is estimation of model fit, where a measure is acquired of how well the model, when path coefficients have been estimated, accords with the available data, using for example a χ^2 test. A high score on such a test is not a confirmation that one has modeled the true causal relations; many other models will always fit the data just as well. Thus, the usual underdetermination of theory by empirical data holds. A good fit is confirmation only of that the model is not significantly undermined by the available evidence.

Causal modeling in terms of diagrams and systems of equations was further developed within economics by for example Trygve Haavelmo (1943), and eventually in the context of social science by Hubert M. Blalock (1964) and O. D. Duncan (1975).

The 1980's saw further important developments in the theory and techniques of causal modeling. The computer scientist Judea Pearl and collaborators introduced the term "Bayesian network" for a causal model where the weights of causal paths are given as probabilistic dependencies under a Bayesian interpretation (e.g. Pearl 1988). Moreover, they developed an algorithm, called *d-separation*, by which independencies between variables can be deduced from the graphical structure of a causal model alone. This method radically simplifies finding testable consequences of causal hypotheses, manually or with the help of a computer. During the 90's, the validity of the d-separation method was proven for an increasing number of kinds of causal models, by Pearl, the philosopher Peter Spirtes, and others (see Spirtes et al. 2000; Pearl's results are collected in Pearl 2009). Specific conditions on causal systems, required for the validity of causal inferences from correlational information, were also identified.

Largely in parallel with structural equation modeling, another school of techniques and interpretations of causal inferences from statistical data, called the Potential Outcomes Framework or the Rubin Causal Model, were developed, mainly by Donald B. Rubin and Paul W. Holland (e.g.: Rubin 1974;

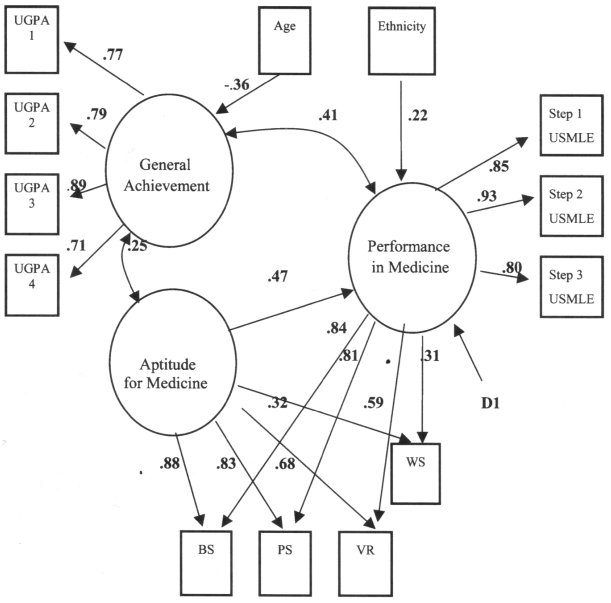

Figure 2. Latent variable path analysis model of UGPA, MCAT, and USMLE (Steps 1–3) latent variables employing ML estimation (n = 24,872). *Note.* Fit indexes: $\chi^2(55) = 11726.28$, $p < .001$ (CFI = .928, RMSEA = .025). UGPA1-4 = Undergraduate GPA Year 1–4; BS = Biological Sciences MCAT Subtest; PS = Physical Sciences MCAT Subtest; VR = Verbal Reasoning MCAT Subtest; WS = Writing Sample MCAT Subtest; Step 1–3 USMLE = United States Medical Licensing Exam Step 1–3.

FIGURE 6.2.2. An example of a fully developed causal diagram, with path coefficients. Ellipses denote latent (unobserved) factors, rectangles measured factors. A bidirectional arrow denotes only correlation. (From Violato and Hecker 2007, used with permission.)

Holland and Rubin 1983; Holland 1986). The Potential Outcomes Framework and causal models share an understanding of causation in terms of counterfactuals involving interventions, but causal diagrams are not usually employed in the former tradition.

Pearl has shown that the Potential Outcomes Framework and structural equation modeling are formally equivalent, in the sense that a theorem in one is a theorem in the other (Pearl 2009, p. 228-31), and has moreover claimed that "SEM provides [...] the formal mathematical basis from which the potential-outcome notation draws its legitimacy" (Pearl 2012, p. 71). The choice between the Potential Outcomes notation and the SEM notation has also been called "a matter of taste" (Elwert 2013, p. 247). This does not, however, mean that they are equivalent research traditions in every practical respect. I focus on the causal modeling tradition that essentially involves diagrams here, because it is the main (although not only) influence on the interventionist approach in the philosophy of causation.

Figure 6.2.2 shows a fully developed causal diagram that includes path coefficients and combines causal paths with both measured and latent variables.

In this model, the set of rectangles with arrows pointing at them from the same elliptical node (e.g., the ones labeled "UGPA") represent measurements that base an estimate of the value of the unobserved factor represented by the ellipsis (i.e., "General Achievement"). The theory depicted concerns primarily the causal relations between these unobserved quantities.

For a brief description of causal modeling in science, that includes the full use of latent variables and a summary of the actual application of the method, see Hox and Bechger 1998. Elwert 2013 provides a more detailed introduction to identifying testable implications from causal models and estimating path coefficients, with a focus on nonparametric models (i.e., models that do not postulate a specific functional form of the dependence between cause and effect, or a particular form of the distributions of variables). Judea Pearl has co-authored two later introductions to causal inference and the understanding of causation in SCM: Pearl et al. 2016; Pearl and MacKenzie 2018. See Cliff 1983 for an older, well-known critical text, and Freedman 2005 for a critical review of linear causal models specifically. What follows here is a simplified presentation of the most central features of a causal model.

6.3. Causal models: graphs and equations

In this section I will stay close to Judea Pearl's theory of causal modeling, which may differ somewhat in notation and interpretation from other presentations of Structural Equation Modeling (Pearl 2009).

A causal diagram is a kind of directed graph that can be used to model causal connections in a system and formulate qualitative causal hypotheses. These are hypotheses about causal laws in systems of the type in questions. I.e., the concern is general causal claims, rather than causal explanations of particular events. The causal laws connect event types in the context of a system, where an event type is the instantiation of some property, or some quantity having a certain magnitude, both denoted in a model by a variable taking on some value. *Nodes* in the graph thus stand for properties or quantities of parts of the system, that may be observed or unobserved. In the context of a causal system, these are often called *factors*. The arrows, or *directed edges*, between nodes stand for direct causal influences. This connection is "direct" only relative to the factors that are modeled—there is no suggestion that these factors are spatiotemporally contiguous or the like.

A *path* is some way of traversing through the graph by following some edges, independently of their direction, and passing any edge at most once. Two nodes are thus connected by a path if one can reach one from the other by following edges in this way. A *directed path* moreover consistently follows the direction of the arrows. A directed path is also sometimes called a *causal path*. If there is an arrow in the model from X to Y, then X is a *parent* of Y and Y a *child* of X. If there is a directed path from X to Y, then X is an *ancestor* of Y, and Y a *descendant* of X. We will restrict ourselves to graphs without cycles (although graphs with cycles can also be treated within the theory). That is, in the graphs we consider it's not possible to return to some node by following a directed path. Such graphs are usually called *DAG*s: directed, acyclic graphs.

Often, rectangular nodes denote observed, or measured, factors in the system, while ellipses denote unobserved factors, and I will follow that convention.

A complete causal model consists of a causal diagram together with a set of corresponding equations. The nodes in the diagram correspond to variables, and the edges to coefficients, in the equations. In the scientific literature, "variable" is often used to denote both a symbol used in an equation, a node in a causal diagram, and the property or quantity in the modeled system that these stand for, and in this chapter and the next I will often be guilty of the same equivocation. Thus we may say that rectangular nodes stand for observed variables and elliptical ones for latent variables.

The diagram (without path coefficients) expresses only a qualitative hypothesis or assumption about what causal connections there are between the variables in the system, while the degree of variation in some variable that is causally due to some other variable is provided by the coefficients in the equations. The purely qualitative hypothesis expressed by the diagram can however have implications for what variables are independent of each other, and these implications can be testable against statistical data. Much of SEM is focused on the *identification* of what coefficients can be estimated based on the available data and given the structure of the causal hypothesis, and the methods of *estimation* for these coefficients. In this chapter, the focus will be on the former.

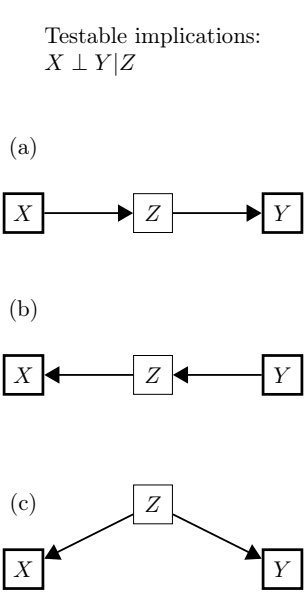

FIGURE 6.3.1. Three causal diagrams implying the same conditional independencies.

By "$X \perp Y$" we mean that variable X is independent of variable Y in the system. These are "population" variables: they have their distributions in the population as a whole or the space of all possible outcomes, not in a particular collected sample. "Independence" means probabilistic independence, that is, X and Y are independent if and only if the probability distribution over Y's possible values is the same regardless of the value of X (i.e., $Pr(Y|X) = Pr(Y)$). This relation is, as we have already observed, symmetric. We assume moreover that if X and Y are observed, then such independence is significantly empirically confirmable in a sufficiently large unbiased sample, and we thereby abstract away from the often considerable practical difficulties of collecting (and knowing that you have collected) such a sample. Thus, if the qualitative causal hypothesis expressed in a causal diagram implies independencies between variables, these are in principle testable consequences of the hypothesis.

The theory of *d-separation* gives us the graphical conditions under which two variables are independent in a causal

model. We will define the d-separation condition later in this section. The general independence notion is that of *conditional independence*. "$X \perp Y|Z$" means that X and Y are independent *conditional on* Z. To condition on Z means, generally, to include information about the value of Z in the analysis. Conditional independence is thus in principle empirically testable, if the variable that needs conditioning on is also part of the data. Below I will start by illustrating the d-separation criteria, and some properties of causal diagrams in general, by a few simple examples. I will describe the corresponding causal equations towards the end of this section.

Figure 6.3.1 shows three different causal hypotheses with respect to the same three observed variables X, Y, and Z. In this and the following diagrams, I have indicated by a bold outline the variables whose relation we are particularly interested in. In the figure, (a), (b), and (c) all have the same testable consequences. This situation illustrates how correlation does not imply causation, and corresponds to the theoretical problems of establishing the direction of causation and excluding spurious correlations, based only on observable regularities, that we have discussed in previous chapters.

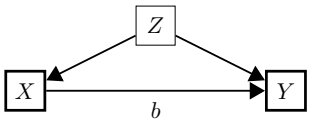

FIGURE 6.3.2. X is a direct cause of Y, and Z is a confounder.

More specifically, in (a), since Z is the direct cause of Y, if we include information about the value of Z, knowing in addition the value of X does not provide any further clue as to the value of Y. Hence, conditioning on Z makes X and Y independent ($X \perp Y|Z$). This obviously works just the same if the causal path between X and Y is reversed, as in (b). Thus, based only on the implied conditional independencies, we cannot tell the difference between (a) and (b). Moreover, in (c) X and Y are correlated only in virtue of sharing Z as a common cause. So, if we know the value of Z, knowing the value of X provides no additional information about the value of Y, and vice versa. Again, X and Y are independent conditional on Z, so that (a), (b), and (c) are in fact empirically indistinguishable, as expected. That is, employing boxes and arrows does not in itself create any new ways of inferring causation from correlations. But since this way of expressing causal relations, and inferring independencies from them, makes formulating more complex causal hypotheses so much easier than

Testable implications:
$X \perp Y$
$X \perp R|Z$
$Y \perp R|Z$

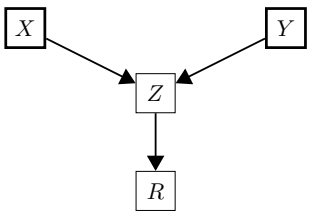

FIGURE 6.3.3. X and Y are independent, but dependent conditional on Z or R.

expressing them in natural language, we can examine more complicated structures and discover what kinds of structures *are* empirically distinguishable, and under what assumptions.

In the model in figure 6.3.2, X is a cause of Y, with the coefficient b, and Z is again a common cause of X and Y. This diagram implies no conditional independencies. However, it does imply that for us to be able to estimate b from our data, we must eliminate the *bias* introduced by Z's influence on X and Y, by conditioning on Z. This is also called *adjustment*, and Z is here a *confounder* relative to b. This is in a sense a generalization from figure 6.3.1 (c), if that diagram is taken to express the special case where $b = 0$. However, we shall see below that under the common assumption that a model is "faithful," it follows that all paths have non-zero coefficients.

Now consider the structure in figure 6.3.3. Here, X and Y are assumed to be causally independent. However, if we were to condition on Z, then X and Y would become dependent, since two independent variables become dependent when we condition on one of their common effects. Imagine that Z here is the number of cookies in a jar, and that X and Y are the number of cookies added or removed from the jar by Max and Kay (and that they are the only ones eating and baking cookies). Max and Kay may eat and bake cookies independently of each other, but if we know how the number of cookies in the jar has changed (i.e., condition on Z), then, if we know how many cookies were added or removed by Max, we can calculate how many cookies were added or removed by Kay. Thus, X and Y are dependent conditional on Z ($X \not\perp Y|Z$). Moreover, as the value of R is an effect of Z (perhaps it's the measured weight of the cookie jar), conditioning on R has the same consequence of making X and Y dependent.

Figure 6.3.3 shows an hypothesis in which X and Y are causally independent of each other. We can generalize, as we did with figure 6.3.1 (c), and add an arrow between X and Y. To then be able to estimate the coefficient for that arrow we would need to *not* condition on Z or any of its descendants, as this would otherwise introduce a confounding bias in the estimate. Such confounding can happen unintentionally, through a misidentification of Z's relations to other factors in the studied system, or unknowingly as a consequence of a biased sampling procedure. This sort of bias has been called "collider bias," among other things (Elwert 2013, p. 250).

We can now state the full d-separation criterion that gives us the conditional independencies implied by a causal model. That they are *conditional* independencies means specifically that d-separation is relative to a set of variables in the model. Let $\mathbb{V}$ be such a set of variables, that we are conditioning on. If a node on some path has two incoming edges (arrows pointing at it), as in the case of Z in figure 6.3.3, we call that node a *collider* relative to any path containing those edges. If a node is not a collider, we call it a non-collider. Then the d-separation condition can be formulated as follows. (For an equivalent formulation, see Pearl 2009, p. 16-17.)

> **D-separation:** In a diagram, two variables X and Y, that are connected by some path, are d-separated relative to $\mathbb{V}$ *if and only if* every path

between them is blocked relative to $\mathbb{V}$.
A path $\mathcal{P}$ between X and Y is blocked relative to $\mathbb{V}$ *if and only if (i)* $\mathcal{P}$ contains a collider n, and neither n nor a descendant of n is in $\mathbb{V}$, or *(ii)* $\mathcal{P}$ contains a non-collider that is in $\mathbb{V}$.

That two variables are d-separated relative to $\mathbb{V}$ then means that they are independent conditional on the variables in $\mathbb{V}$. A special case is when two variables are independent relative to the empty set $\mathbb{V}$, in which case they are unconditionally independent. X and Y in figure 6.3.3 provide an example. In general, the test implies conditional independencies. Consequently, conditioning only on Z or R in figure 6.3.3 unblocks the path between X and Y—since this violates condition *(i)* and condition *(ii)* is then also not satisfied—thereby making X and Y dependent given any $\mathbb{V}$ that includes Z or R but no variables such that condition *(ii)* would be satisfied. I.e., the model implies X and Y are dependent conditional on Z or R. Conditioning on Z in any of the diagrams in figure 6.3.1 instead blocks the path between X and Y, by satisfying condition *(ii)*—note that Z is a non-collider in each of those diagrams—thus making X and Y independent conditional on Z ($X \perp Y|Z$).

We might note here that if two variables are independent relative to *every* possible $\mathbb{V}$, then it follows that they are disconnected in the graph—that there is no path between them—since every path between X and Y either has or doesn't have a collider on it, and if it does then X and Y are dependent conditional on that variable, while if it doesn't then they are unconditionally dependent. Since two disconnected variables are obviously independent for any $\mathbb{V}$, it follows that X and Y are independent for every $\mathbb{V}$ if and only if they are disconnected in the graph.

As an example of a slightly more complicated structure, that may introduce an interesting kind of error in estimation, consider figure 6.3.4. The previous causal diagrams have contained no information about the temporal ordering of the events in the system. Such information is often not available in the contexts of social science and econometrics, since events tend to be slow, with ill-defined start- and endpoints. Here I have nevertheless introduced a time axis just to indicate this temporal ordering. R and S are unobserved variables, as indicated by their nodes being elliptical. Note that there are no testable independencies implied by this model, due to the fact that the variables that would require conditioning on (R and S) are unobserved. If figure 6.3.4 depicts the true causal relations in the system, then Z precedes X and Y, and is correlated with both. Based only on this, we might be lead to think that Z is a common cause of X and Y, and that we therefore need to adjust for Z when estimating b. However, conditioning on Z in this system *introduces*—rather

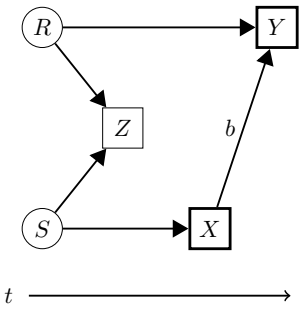

FIGURE 6.3.4. Conditioning on Z introduces bias in estimating b.

than eliminates—bias in our estimate. This is a variant of collider bias, sometimes called "M-bias." If we have conditioned on Z in this situation, we need to also adjust for either R or S to eliminate the bias we thereby introduced. (This blocks again the path between X and Y that passes through Z.) But this is not possible if R and S are unobserved. Note that in a structure like this one, but in which X and Y are not causally related, in a sample that is inadvertently biased with respect to Z, X and Y may be seen to correlate, even though neither is a cause of the other, and they don't share a common cause. This is not in contradiction with the Common Cause Principle (see the first chapter, and below), since it is a case of conditional dependence between X and Y, where we are unaware of having conditioned on Z. Nevertheless, it might be hard or impossible to tell the difference in a real situation.

We now need to explain two conditions on causal models, for making these causal inferences, that I have so far assumed to be satisfied. First, the *Causal Markov Condition* is necessary for the validity of d-separation, and thus at the very core of causal inferences in this theory (Spirtes et al. 2000, sect. 3.4.1, Pearl 2009, sect. 1.2 and 1.4.2).

> **Causal Markov Condition:** A causal model $\mathcal{M}$ with diagram $\mathcal{D}$ satisfies the Causal Markov Condition *if and only if* every variable $\mathcal{V}$ in $\mathcal{M}$ is independent of its non-descendants in $\mathcal{D}$ conditional on its parents in $\mathcal{D}$.

In brief, the Causal Markov Condition states that the full set of direct causes of an event screens off that event from everything that isn't a direct or indirect effect of the event. This requires some explanation. Consider two applications to the literally straightforward diagram (a) in figure 6.3.1. First, to assume that the model to which (a) belongs satisfies the Causal Markov Condition implies that $X \perp Y|Z$. This is what the d-separation test delivers. But, second, it is *not* implied that $Z \perp Y|X$. How can this be? Just looking at the diagram we might think that, assuming linearity for simplicity, X determines Z with some coefficient b_{ZX}—i.e. $Z = b_{ZX}X$—and Z determines Y with some coefficient b_{YZ}—i.e. $Y = b_{YZ}Z$. But if this is the case, then $Y = b_{YZ}b_{ZX}X$, and knowing the value of Z adds no information about the value of Y once we know the value of X. Hence, X screens off Z from Y. We encountered this situation already in connection with Reichenbach's probabilistic theory of causation and the screening off condition (chapter 1). There is thus, here as there, a necessary assumption, for the theory to work, that the parents of a variable do *not* determine the value of that variable, but that there are additional, implicit factors that also affect it. These are not usually drawn into the causal diagram, but they can be, and figure 6.3.5 has been amended with these implicit factors.

Every variable $\mathcal{V}$ in a causal model is assumed to implicitly be connected by an incoming arrow to a $U_{\mathcal{V}}$, that is normally not drawn in the diagram. Each $U_{\mathcal{V}}$ is assumed to be *exogenous* to the system. A variable is exogenous if and only if it has no incoming arrows, i,e. no causes in the system. A variable that is not exogenous is *endogenous*. Each $U_{\mathcal{V}}$ is also unobserved, and independent of every other $U_{\mathcal{W}}$. $U_{\mathcal{V}}$ is understood, then, as summarizing all

the independent causal influences on $\mathcal{V}$ that are not in the system (i.e., not part of the model), but rather enters from the environment of the system. It is in other words assumed that every measured factor in a system is affected by some unknown causes that are independent of those corresponding influences on the other known factors.

This is not just an assumption that accommodates the theoretically useful Causal Markov Condition and thereby the d-separation algorithm, it is also supported by observation: whenever we try to estimate a path coefficient, for example through regression analysis, the observed values of the outcome variable will deviate somewhat from that dictated by the independent variable(s) and best fitted coefficient(s). This deviation is called the *residual* and usually denoted by "u" in the equation that estimates the dependence between the variables. (In principle, we can fit a line that goes exactly through the datapoints, and therefore has zero residuals, by using a sufficiently complex polynomial to define the line, but this "overfitted" line will tend to not be a good fit for *future* observations of this system, which suggests that the residuals are indeed due to further, unobserved and independent factors affecting the outcome.)

Testable implications:
$X \perp Y|Z$

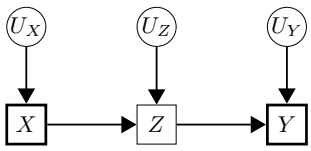

FIGURE 6.3.5. Diagram including noise variables U_X, U_Z, U_Y.

As $U_{\mathcal{V}}$ is unobserved by definition, we can only know about its influence on $\mathcal{V}$ by way of these residuals. This deviation is often called "error" or—especially under a causal interpretation—"noise," and we may call $U_{\mathcal{V}}$ a "noise variable." While the residuals are the only way for us to estimate $U_{\mathcal{V}}$, it's important to note that they are not the same thing. Generally, residuals may be due to measurement error, and they may be correlated with an independent variable or with each other. $U_{\mathcal{V}}$, on the other hand, by stipulation stands for those exogenous influences that affect $\mathcal{V}$ alone.

Testable implications:
$X \perp R|Z$
$Y \perp R|Z$

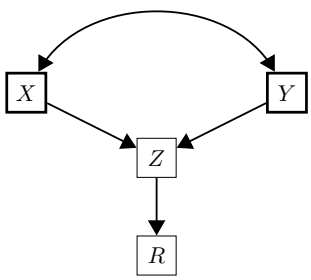

FIGURE 6.3.6. X and Y are hypothesized to have an unnamed common cause that explains their correlation.

Again, since $U_{\mathcal{V}}$ is assumed to be unobserved, we can only treat it as a random variable, the distribution of which may be estimable from the observed residuals. $U_{\mathcal{V}}$ does by assumption have a mean value of zero, and some positive variance. In linear models it's common to assume $U_{\mathcal{V}}$ to be normally distributed, but this is not a requirement in causal models. That $U_{\mathcal{V}}$ is a random variable means that the outcome

variable $\mathcal{V}$, too, is random, and that fixing the values of $\mathcal{V}$'s causes only affects the probability distribution over the possible outcomes of $\mathcal{V}$, it does not fix the value of $\mathcal{V}$. Any causal model within this framework is thus strictly epistemically indeterministic (called "pseudoindeterministic" by Spirtes et al. 2000, sect. 2.5). A truly indeterministic system could be modeled the same way, but the residuals, and hence the $U_{\mathcal{V}}$s, would then not have the interpretation they have been given here. Moreover, as $U_{\mathcal{V}}$ is unobserved, we can by assumption not condition on it, and it doesn't count among $\mathcal{V}$'s parents. Therefore, conditioning on $\mathcal{V}$'s parents does not fix the value of $\mathcal{V}$, nor the values of its descendants, which explains why, under these assumptions, it is not the case that $Z \perp Y|X$ in figure 6.3.5.

Pearl observes that the Causal Markov Condition follows from two assumptions about causation: *(i)* Reichenbach's Common Cause Principle, which states that if two variables are correlated, then one is the cause of the other or they have a common cause, and *(ii)* if two variables that are in the model have a common cause, then that cause is also in the model (Pearl 2009, p. 30).

We can also prove that the Common Cause Principle follows from the Causal Markov Condition (Arntzenius 2010, note 4). In virtue of *(ii)*, it is assumed that all exogenous variables in a model are independent. Conversely, if two variables are not independent but by hypotheses are not related as cause and effect, then they must be connected to a common cause. Notation-wise, if there is no observed variable that fits the bill, then these variables may be connected to a hypothetical latent variable, or they may be connected just by a bidirectional arrow, as in figure 6.3.6. In this diagram, d-separation does not imply that $X \perp Y$, since the bidirectional arrow is an unblocked (and unblockable) path between X and Y. A model in which any common cause of two variables in the model is also in the model is called *causally sufficient* by Spirtes et al. (2000, p. 22).

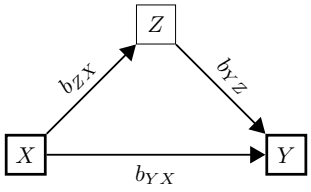

FIGURE 6.3.7. X is both a direct and an indirect cause of Y.

Next, if the *faithfulness* condition is satisfied, then d-separation gives us *all* the conditional independencies in the system—if not, then there may be further independencies. Figure 6.3.7 shows a system in which X causally influences Y along two different paths. If the causal influences along these different paths are such as to cancel out, then although X is a cause of Y, X and Y are nevertheless unconditionally independent in the system. More specifically, if we assume that the causal dependencies are linear, then if $b_{YR}b_{RX} = -b_{YX}$, then X and Y are unconditionally independent, but this is not given by the d-separation test. In general, the faithfulness condition states just that all conditional independencies in the system are given by the d-separation test applied to its model (Spirtes et al. 2000, p. 13).

That a system satisfies faithfulness is a contingent matter, and violations are clearly within the realm of the physically possible, but it has been argued that they are *a priori* improbable. (Specifically, that in the space of linear dependencies they have *a priori* zero probability (Spirtes et al. 2000, p. 41).) Faithfulness may therefore seem to often be a fair assumption. But the assumption of faithfulness can also be given a more principled character. It may be regarded as the assumption that "all independencies are structural," as opposed to contingent on the precise values of certain parameters such as the bs in figure 6.3.7 (Pearl 2009, p. 49).

We have briefly reviewed causal diagrams, what probabilistic independencies can be inferred from the structures they depict, and what conditions must hold on the parameters and joint probability distributions of the variables for these inferences to be valid. Let's finally look at the system of equations that, together with a diagram, makes up a causal model. In table 6.3.1 , the dependencies between the variables that are stated in figure 6.3.7 are assumed to be linear. This is a special, but common, case, and means that the dependent variable $\mathcal{V}$ is a linear function of the independent variables and the random noise $U_{\mathcal{V}}$. $U_{\mathcal{V}}$ is also commonly assumed to be normally distributed. If none of these assumptions are made about forms of functions and distributions, then the model is called "nonparametric," and the general expression of a causal equation is then "$\mathcal{V} \stackrel{c}{=} f(PAR(\mathcal{V}), U_{\mathcal{V}})$," where $PAR(\mathcal{V})$ is the set of parents of $\mathcal{V}$. Another form of nonparametric causal model, most strongly associated with Pearl's treatment, states dependencies as conditional probabilities only. A causal model then consists in a joint probability distribution P for a set of variables, and a causal diagram, or graph, G, representing a causal structure over these variables. I will focus on the functional representation of causal models here.

In general, a system of structural equations has one equation for every node $\mathcal{V}$ in the corresponding diagram, and in each such equation, one term for every node that $\mathcal{V}$ is connected to with an incoming arrow, including the implicit noise term. For short, I will call the equation in which $\mathcal{V}$ is the dependent variable (i.e. on the left hand side) "$\mathcal{V}$'s equation." The system of structural equations in a causal model moreover has two special and vital properties. Firstly, by "$\stackrel{c}{=}$" I mean to indicate that these equations have an explicit causal interpretation. That means in particular that they are asymmetric. I.e., while equation 2 implies that $X = \frac{1}{b_{ZX}} Z - U_Z$, it does *not* imply that $X \stackrel{c}{=} \frac{1}{b_{ZX}} Z - U_Z$, as this latter expression states that Z is a cause of X. That is, these structural equations do not state merely mathematical equalities, but something *physical* about the relations between the factors in the system. Secondly, each equation denotes an autonomous causal mechanism in the system, in the sense that it could be modified while the parts of the system denoted by the other equations remain

1. $X \stackrel{c}{=} U_X$
2. $Z \stackrel{c}{=} b_{ZX} X + U_Z$
3. $Y \stackrel{c}{=} b_{YX} X + b_{YZ} Z + U_Y$

TABLE 6.3.1. Linear structural equations specifying the causal relations in figure 6.3.7.

the same. This is usually called "modularity," and is required for the interpretation in this theory, of the causal arrow in terms of possible interventions, to which we turn next.

6.4. The causal arrow, intervention, and modularity

In this section, I will discuss two aspects of how we understand the arrows in a causal diagram, that are not always distinguished very explicitly in the scientific method literature. There is, firstly, the question of what causal claims *mean*. This question may be more or less overtly a question of *operational meaning*. As such, it is closely connected to the ability of causal models to *predict*, for a certain causal system, the outcomes of interventions, such as experiments or policies. Secondly, there is the question of what, if anything, it is *in the world* that makes these causal claims true. This question is rarely addressed, but of obvious philosophical interest.

In the history of the statistical sciences, the meaning of causal claims has sometimes been cashed out in terms of idealized or hypothetical experiments. For example, in his "Actions, Consequences, and Causal Relations," the econometrician Guy Orcutt presented an interpretation of causal claims within his science in terms of the consequences of actions, and claimed that "the statement that z_1 is in a causal chain leading up to z_2, or that z_1 is a cause of z_2, is just a convenient way of saying that if you pick an action which controls z_1, you will also have an action which controls z_2" (Orcutt 1952, p. 307). When framing this interpretation, Orcutt assumed an "idealized experiment," in which the experimenter is "dealing with variables that he alone acts upon" (Orcutt 1952, p. 306). Orcutt employed a kind of causal graphs, and his main concern in this paper was "the inference of causal relations suitable for specifying the consequences expected from action," e.g., of policy decisions (Orcutt 1952, p. 305). These features are indicative also of later works in causal modeling.

Moreover, in the Potential Outcomes Framework for causal inference mentioned above, a causal effect is defined relative to two possible, mutually exclusive causes of a single outcome variable in a system, and the difference in the outcome between these. In "Statistics and Causal Inference" the statistician Paul W. Holland, one of the main proponents of this theory, names these two causes symbolically "treatment" and "control." The use of experimental vocabulary is however more than a convenience. Although Holland states that "It is not that I believe an experiment is the *only* proper setting for discussing causality, but I do feel that an experiment is the *simplest* such setting" (Holland 1986, p. 946), he nevertheless arrives at a motto coined by himself and Donald Rubin: *no causation without manipulation*. He takes this statement to be a consequence of the definition of an effect in the theory, which implies that an outcome under both a treatment and a control need to be definable in principle in the model, for there to be a causal effect. That this restricts what can be causes (i.e., nothing that cannot be manipulated in principle), he takes as a benefit of the theory, in virtue of making the cause concept more specific (Holland 1986, p. 959).

Judea Pearl, in turn, characterizes a principal difference between causal modeling and classical statistical methods by stating that the causal modeling theory allows predictions, not only of events in systems that are passively observed, but also of the consequences of certain kinds of *changes* into systems—i.e. manipulations, or more generally interventions:

> Standard statistical analysis, typified by regression, estimation, and hypothesis-testing techniques, aims to assess parameters of a static distribution from samples drawn from that distribution. With the help of such parameters, one can infer associations among variables, estimate the likelihood of past and future events, as well as update the likelihood of events in light of new evidence or new measurements. These tasks are managed well by standard statistical analysis so long as experimental conditions remain the same. Causal analysis [by way of causal modeling] goes one step further; its aim is to infer not only the likelihood of events under static conditions, but also the dynamics of events under *changing conditions*, for example, changes induced by treatments or external interventions, or by new policies or new experimental designs. (Pearl 2009, p. 332.)

Causal analysis can do this because it models not only associations, but the causal laws that governs the system. Spirtes et al.—who were the first to apply these sorts of causal models to the specific problem of predicting the effects of interventions and to provide an explicit causal interpretation of this—state on that same note that the joint probability distribution estimated from observing a system cannot alone be used to predict the probabilities of events in that system under some manipulation of it, but that this probability distribution together with the system's causal structure can (Spirtes et al. 2000, sect. 3.7.2). Thus, the theory connects in a direct way to the main reason for preserving the cause concept that we have encountered in this part of the monograph: the fact that causation is required for distinguishing "effective strategies" from ineffective ones.

This is how to infer the consequences of an intervention into a system. In terms of the causal diagram, an intervention on a variable X that sets X to some value k, is modeled by deleting from the causal model all arrows pointing at X, and assuming that $X = k$. In terms of the model's structural equations, the operation is performed by replacing X's equation by an equation "$X \stackrel{c}{=} k$." That is, an intervention on X in a system $\mathcal{S}$ is here understood as a "remodeling" of the causal structure of $\mathcal{S}$, such that in the new model $\mathcal{S}_I$, X has only a single, implicit cause, that is moreover exogenous to the system—namely the intervention event—and this transformation is accomplished by severing all of X's connections to any causal influences on it from within the system, that it used to have in $\mathcal{S}$ (Pearl 2009, sect. 131). The effect of this intervention on a variable Y is now calculated by solving the equation for Y, and the equations for all the independent variables in Y's equation, in $\mathcal{S}_I$. If Y depends on X in $\mathcal{S}_I$, then we can conclude that X is a cause of Y in $\mathcal{S}$.

Modeling an intervention on X in a system $\mathcal{S}$ then amounts to transforming the model for $\mathcal{S}$ into a new model $\mathcal{S}_I$ according to specific rules, and solving the relevant equations in this model. For it to be possible to perform this transformation—for there to *be* an intervention on X in $\mathcal{S}$—$\mathcal{S}$ must be *modular*. That is, it must be possible to replace the causal mechanism that determines the value of X in $\mathcal{S}$ with one that sets X exogenously, without disturbing the rest of $\mathcal{S}$. If it is *not* possible to intervene on X without changing the other causal mechanisms of the system, then we cannot conclude anything about casual effects in $\mathcal{S}$ from our calculations on $\mathcal{S}_I$. Pearl calls the modification of $\mathcal{S}$ under these constraints a "surgical" operation.

Pearl proceeds to *define* the effect of X on Y in a system in terms of the consequences of an intervention on X. Pearl made the following statement in 1996, in an informal and useful presentation of his theory (reprinted in Pearl 2009):

> We now see how this model of intervention leads to a formal definition of causation: "Y is a cause of Z if we can change Z by manipulating Y, namely, if after surgically removing the equation for Y, the solution for Z will depend on the new value we substitute for Y." (Pearl 2009, p. 417.)

The difference between inferring, in the way described here, the effect of an intervention that sets X to k, and inferring the consequences of $X = k$ in the *unmanipulated* system $\mathcal{S}$ is central, and corresponds to the difference between setting X to k and merely *observing* that $X = k$. From observing that $X = k$ we may be able to infer something about X's *causes* in the system. By setting X to k by an intervention, on the other hand, we by assumption leave X's causes unaffected. Correspondingly, if we observe that $X = k$, then our causal hypothesis may allow us to predict a certain value l of Y with some probability, even if the information that the value of X gives us about the probability that $Y = l$ is a consequence of X and Y having a common cause. But setting $X = k$ through an intervention, by the assumption that interventions are events exogenous to the system, breaks X's connections to any common causes it might have shared with Y, so that this operation only tells us something about Y if X is in fact a cause of Y.

Pearl indicates the difference between the probability of $Y = l$ conditional on the *observation* that $X = k$, and the probability of $Y = l$ conditional on *setting* X to k through an intervention by his use of the $do(\cdot)$ operator. The standard expression for conditional probability "$Pr(Y = l|X = k)$" denotes the former, and "$Pr(Y = l|do(X = k))$" denotes the latter.

Above, I mentioned that one of the principal activities in causal modeling is *identification*, by which is meant the identification of those causal effects—that is, those path coefficients—that can be estimated given a certain causal structure and the available data. Of particular interest are cases where the data are purely observational. The rules of Pearl's "*do* calculus" turns this task into a formal procedure (Pearl 2009, sect. 3.4). If an expression of the form "$Pr(Y = l|do(X = k))$" can be transformed into an expression of the standard form "$Pr(Y = l|X = k)$" by some finite application of the three rules of the *do*

calculus, then the effect of X on Y in the system can be estimated (assuming, again, a sufficiently large unbiased sample of the relevant variables). Thus, we see that the effect of X on Y in a system is again defined in the theory as the change in Y that would occur under an intervention on X. (Pearl's definition of the causal effect of X on Y, which he denotes by "$Pr(Y = l|do(X = k))$," is in 2009, p. 70.)

Finally, Pearl provides an *operational* definition of "structural equation," that he suggests provides the meaning of the term:

> What then *is* the meaning of a structural [path] coefficient? Or a structural equation? Or an error [noise] term? The interventional interpretation of causal effects, when coupled with the $do(x)$ notation, provides simple answers to these questions. The answers explicate the operational meaning of structural equations and thus should end, I hope, an era of controversy and confusion regarding these entities. (Pearl 2009, p. 160.)

The controversy Pearl has in mind regards, I think, the interpretation of these things in the scientific community, rather than in philosophy. The notation in the definition below has been slightly modified for the sake of consistency with the rest of this text (see Pearl 2009, p. 160 for the original formulation).

> **Pearl's operational definition of "Structural Equation":** An equation $Y = bX + U$ is said to be *structural* if it is to be interpreted as follows: In an ideal experiment where we control X to a value k and any other set $\mathbb{Z}$ of variables (not containing X or Y) to a corresponding set of values $\mathbb{L}$, the value of Y is given by $bk + U$, where the value of U is not a function of the values of X and $\mathbb{Z}$.

Since X's effect on Y in $\mathcal{S}$ is determined by how Y *would* change under an intervention on X, independently of any actual such intervention, causal modeling is, from one perspective, a counterfactual theory of causation, and Pearl recognizes this.

For a definition of causation in the terms of a counterfactual conditional to work, it must be that, if X and Y are dependent due to a common cause Z, but X is not a cause of Y, then "If X had been different then (the probability of) Y had been different" is false, since it is the truth or falsity of this counterfactual that determines whether X is a cause of Y or not. This counterfactual is *not* false if our evaluation of it involves a "backtracking" inference from the value of X, to the value of the common cause Z, and then to the value of Y. This procedure corresponds rather to our prediction of the value of Y given a mere *observation* of the value of X, and its result does not depend on whether X is a cause of Y or if they are dependent for some other reason. Thus, it is the assumption that it is the change in Y under a possible, but non-actual, *intervention* on X that prevents a backtracking inference—literally an inference along causal arrows in the wrong direction, from effect to cause. The exogeneity of interventions in the theory—the severing of X's connections to its causes in $\mathcal{S}$—does the same job in this respect, then, as the "small miracles" that prevent backtracking inferences in David Lewis's counterfactual theory of causation (Lewis 1987b).

Pearl dedicates chapter 7 of *Causality* (2009) to an analysis and interpretation of counterfactuals within the framework of his theory, and compares it to Lewis's similarity-based account. He there concludes:

> In contrast with Lewis's theory, counterfactuals are [here] not based on an abstract notion of similarity among hypothetical worlds; instead, they rest directly on the mechanisms (or "laws," to be fancy) that produce those worlds and on the invariant properties of those mechanisms. Lewis's elusive "miracles" are replaced by principled minisurgeries, $do(X = x)$, which represent the minimal change (to a model) necessary for establishing the antecedent $X = x$ [...]. Thus, similarities and priorities—if they are ever needed—may be read into the $do(\cdot)$ operator as an afterthought [...] but they are not basic to the analysis. (Pearl, 2009, p. 239.)

However, there are further important differences between Pearl's and Lewis's use of counterfactuals, in their respective theories. It is clear that deriving a similarity relation from a causal structure, as Pearl proposes, is the *inverse* of what Lewis tried to accomplish in defining causation in terms of counterfactual conditionals, and evaluating these in turn based on the overall relative similarity ordering of possible worlds. Lewis was after a reductive analysis of causation that conformed to his Humean concerns. It seems fair to say that Pearl takes the counterfactuals to be implied by the causal structure, rather than the other way around. More on Pearl's causal primitivism below.

As X's effect on some other variable in the system $\mathcal{S}$ depends on there being a possible intervention on X, and the possibility of an intervention in turn depends on the *modularity* of $\mathcal{S}$, it is a necessary condition for something to be a cause that the system in which it is a cause is modular with respect to that factor. The requirement that all systems are modular with respect to their causes can, in a way, be regarded as an interventionist addition to the unmanipulable causes problem (see the previous chapter). This implication has also been criticized in particular by Nancy Cartwright. She has proposed that many causal systems are not modular, and described what she takes to be a common sort of counterexample: the carburetor of a car engine (Cartwright 2007, p. 15-16). In brief, Cartwright suggests that in a model of this causal system, the amount of gas in the combustion chamber is determined by several factors, each relying to some extent, for their degree of influence, on the geometry of the chamber. Thus, changing the mechanism of one of these factors entails modifying the geometry of the combustion chamber, represented by a parameter which also occurs in other equations in the model, thus violating modularity. Pearl has responded to this in 2009 (sect. 11.4.7), where he proposes, on the one hand, that it is in general sufficient that a *symbolic* intervention can be performed on the causal model, for the determination of causal effects, and on the other hand that we nevertheless could isolate the individual causal contributions in this particular example. This problem acquires a more general form in Woodward's treatment, where the connection to the traditional problem of unmanipulable causes is also clearer.

It is tempting—to philosophers at least—to equate claims in this literature, about the meaning of causal claims being given by claims about what would happen under a hypothetical intervention—or an explicit definition of causation to the same effect—with that same claim as it would be interpreted in a philosophical context. That is to say, such a claim would normally be understood there as giving the *truth conditions* of said causal claims. It is generally hard to know whether any such beliefs are involved in the scientific context. However, Pearl in particular has denied, in increasingly explicit terms, that this is what is intended. Instead, he has spoken of causal influences as they exist in the world in ways that suggest causal primitivism. He has recently liked to describe a factor Y, that is causally dependent on another factor X, as "listening" to X and determining "its value in response to what it hears" (Pearl and MacKenzie 2018, p. 13). This formulation suggests to me that it is the fact that Y is "listening" to X that *explains* why and how Y changes under an intervention on X. That is, what a possible intervention does, is to isolate the influence that X has on Y, in virtue of Y's "listening" to X. Thus, Pearl's theory does not imply an interventionist theory of causation, as we understand that concept in this monograph. This, moreover, suggests that the intervention that is always available, for any cause that is represented by a variable in a causal model, is a *formal operation*. I take this to be supported by the way he responds to Nancy Cartwright's objection that modularity does not hold of all causal systems: it is sufficient that a symbolic intervention can be performed. Thus, the operation alluded to in Pearl's operationalization of causation is a formal operation, always available, regardless of whether it corresponds to any possible intervention event or not. (A question that, from the formal perspective, and when causal relations are taken as theoretical primitives, moreover appears beside the point.)

CHAPTER 7

James Woodward: Interventionism

7.1. Introduction and meta-theory

7.1.1. Causation and causal explanation. James Woodward's *Making Things Happen* (2003) is easily one of the most influential books in the philosophy of causation over the last several decades. Its subtitle is "A Theory of Causal Explanation"—but my focus will be on the view of causation proper that Woodward presents there. I will thus not discuss Woodward's theory of (causal) explanation as such. That theory revolves around two central notions. One is that causal explanation provides information that is potentially relevant for *manipulation and control* (Woodward 2003, p. 10). The other is that the *depth* of a causal explanation has to do with the degree of *invariance* in the causal connection that provides the explanation, a causal connection being more invariant if it holds under a wider range of variable values and background conditions (Woodward 2003, ch. 6, examples p. 260-261). For a discussion of Woodward's theory of explanation specifically, with replies by Woodward, see Elliot Sober's contribution, and Woodward's response, in Humphreys et al. 2006.

In this chapter I will focus on Woodward's presentation of his theory and some critical responses to it. I will look closer at the logic of the theory, and some implications for the goals Woodward commits to in *Making Things Happen*, in the next chapter.

7.1.2. The meaning and function of causal claims. Woodward builds his treatment on the earlier theoretical work done mainly by Judea Pearl and Peter Spirtes, Clark Glymour, and Richard Scheines, that I outlined in the previous chapter, and he also references Christopher Hitchcock (e.g. Hitchcock 2001a). Woodward describes the particular goal of his work in the following way.

> As I understand Pearl's enterprise, it takes as primitive various qualitative notions of causal dependence (e.g., the notion of X being directly causally relevant to Y), defines the notion of an intervention by reference to this notion, and then shows us how to calculate or estimate various quantitative causal notions (such as the magnitude of the total effect of X on Y) in terms of this framework. Spirtes et al. are, by their own account, less interested in issues about what various sorts of causal claims mean and focus instead on problems of causal inference or discovery from statistical data. By contrast, I

> have nothing to say about issues having to do with calculating quantitative magnitudes, estimation, identifiability, or causal inference. Instead, my enterprise is, roughly, to provide an account of the meaning or content of just those qualitative causal notions that Pearl (and perhaps Spirtes et al.) take as primitive. (Woodward 2003, p. 38.)

Woodward's goal, then, is the familiar-sounding one of providing a theory of the content of causal claims, and he also describes this project in terms of "providing truth conditions for claims" employing a variety of causal concepts (2003, p. 95). But Woodward is also keen to point out that this is not a matter of a traditional conceptual analysis, mainly because his project does not aim merely to give an interpretation of actual language use, but to be "*revisionary* or *normative*" as well (2003, p. 7). That is, he intends to say something also about what we *ought* to mean by "X is a cause of Y," given the practical and theoretical reasons we have for employing causal language in the first place. This may be seen as a contrast to some earlier authors in the philosophy of causation, whose goals have been mainly descriptive. David Lewis, for example, says:

> When common sense delivers a firm and uncontroversial answer [to a causal question] about a not-too-far-fetched case, theory had better agree. If an analysis of causation does not deliver the common-sense answer, that is bad trouble. (Lewis 1987b, appendix E.)

But, on the other hand, Woodward's project seems closely related to that called "*explication*" by Rudolph Carnap (as also noted by Michael Strevens (2008, p. 184)). It is similar at least in that the explicatum should be descriptively adequate with respect to the everyday concept, but does not need to exactly match it—thus the project can be revisionary—and also in the requirement that the resulting concept should be "fruitful," although Woodward would surely understand this in more straightforwardly practical terms than Carnap did (Carnap 1962, chapter 1).

In later writings Woodward has further distanced himself from what he takes to be the common methods in modern philosophy of causation, calling his own approach "functional." He contrasts the functional approach to three other types of projects found in the philosophy of causation, that he call "the *metaphysical project*, the *description of ordinary usage project*, and the *how does causation fit with physics project*" (Woodward 2014b, p. 692). He characterizes the functional approach in the following way.

> I have in mind an approach that takes as its point of departure the idea that causal information and reasoning are sometimes useful or functional in the sense of serving various goals and purposes that we have. It then proceeds by *trying to understand and evaluate various forms of causal cognition in terms of how well they conduce to the achievement of these purposes.* Causal cognition is thus seen as a kind of epistemic technology—as a tool—and, like other technologies, judged in

> terms of how well it serves our goals and purposes. (Woodward 2014b, p. 693-694. Emphasis added.)

Woodward lists some issues especially appropriate to address within the functional project, such as what useful distinctions we can make between different kinds of causal claims, identifying the preconditions of useful causal thinking in organisms, and the conditions for reliable causal inference from empirical evidence. In *Making Things Happen* as well as in later writings, he in particular emphasizes how an interventionist view of causation can guide formulations of causal claims, and make them more precise, by connecting them to hypothetical, idealized experiments, that may or may not be possible to execute. For example, we may make claims about how an individual's *species*, *gender*, or *age*, is a cause of something—but if we have no clear notion of how these properties could be changed in some particular individual by an intervention, then our causal claim is unclear and may require re-specification. And if it proves to be the case that something modeled as a variable cannot, for metaphysical, logical, or conceptual reasons, be changed by an intervention, then we ought to conclude that this variable in fact cannot be a cause of anything. Moreover, a causal claim may be analyzed in different ways, appealing to different possible interventions, and this exercise may again be clarificatory. Woodward uses an example taken from Paul Holland, where the claim is that "Being female causes one to be discriminated against in hiring and/or salary" (Woodward 2003, p. 115). Woodward claims, like Holland, that what precisely is meant to be the cause in this claim is unclear, and that it is helpful to consider what it is we imagine that we would want to manipulate to show this claim to be true. Different possibilities would be the actual gender of a person, or the *belief* about the gender of that person that is held by a potential employer. Clearly, these are the objects of very different manipulations.

As described now, Woodward's project appears to be about understanding the nature of, and conditions for, reliable and useful *causal cognition*, and in recent writing Woodward even seems to contrast his current goals with that of finding truth conditions for causal claims, which he there associates rather with a kind of metaphysics he has no interest in (Woodward 2017).

Nonetheless, Woodward clearly states that the theory presented in *Making Things Happen* is a product of the functional approach, and that theory consists largely in what looks like necessary and sufficient truth conditions for several different kinds of causal claims, and they are sometimes treated and defended that way by Woodward (see section 7.2.2, below), so we might conclude that this is at least one form that a functional theory of causal cognition may take. In its focus on the practical utility of philosophical theories of causation, there is a definite continuity between *Making Things Happen* and these later statements, but there also appears to be a move in focus, away from a philosophically traditional semantic project, toward something closer to what Pearl and Spirtes et al. had been doing in their works. I will assume that what Woodward says in *Making Things Happen* stands, in this review of his interventionist theory of causation.

7.1.3. Manipulationism and realism. In Woodward's theory "intervention" is a technical term inherited from the literature on causal modeling, mainly Pearl's, but he frequently calls the theory itself "manipulationist." He outlines his manipulationist perspective on causation thus: "On this way of looking at matters, our interest in causal relationships and explanation initially grows out of a highly practical interest human beings have in manipulation and control; it is then extended to contexts in which manipulation is no longer a practical possibility" (Woodward 2003, p. 10). He further associates this perspective with the tradition, found in some parts of the statistical sciences, of clarifying causal claims by reference to the consequences of hypothetical, idealized experiments, that we saw some examples of in the previous chapter.

Thus embracing a manipulationist view of causation, Woodward also emphasizes differences between his efforts and those of the earlier causal manipulationists that we have reviewed in previous chapters. Mainly, he takes the difference to be about reductionist ambitions.

> Philosophical defenders of the manipulability conception have typically attempted to turn the connection between causation and manipulability into a reductive analysis: their strategy has been to take as primitive the notion of manipulation (or some related notion like agency or bringing about an outcome as a result of a free action), to argue that this notion is not itself causal (or at least does not presuppose all of the features of causality we are trying to analyze), and to then attempt to use this notion to construct a noncircular reductive definition of what it is for a relationship to be causal. Philosophical critics have (quite reasonably) assessed such approaches in terms of this aspiration (i.e., they have tended to think that manipulability accounts are of interest only insofar as they lead to a noncircular analysis of causal claims) and have found the claim of a successful reduction unconvincing. [...] I agree with the philosophical critics that the reductionist version of the manipulability theory is unsuccessful. (Woodward 2003, p. 27-28.)

Woodward further explicitly recognizes a relationship between anthropocentrism and reduction: attempting to *ground* causation in free action or agency will make for a subjective and anthropocentric notion of "cause," and the only way for a manipulation theory to steer clear of such a consequence is to abandon hopes of reduction. Thus: "[I]t is crucial to my argument that an account of causation and explanation can be worthwhile and illuminating without being reductive" (2003, p. 21).

In his criticism of Peter Menzies's and Huw Price's manipulationist theory, Woodward highlights his motive for avoiding these anthropocentric implications of a reductivist manipulation theory. He there argues that any adequate account of causation must be *realist*, against especially Price's antirealism or quasi-realism. I outlined Woodward's argument in section 5.5.3, above. In the final section of this chapter I will suggest that Woodward's realism is the more

illuminating difference between him and other manipulationists, although it is also closely related to the issues of reduction and circularity. I also mentioned in that section that Woodward recognizes one source of relativity in causal judgments, that have to do with judgments about "serious possibilities." This appears closely related to Price's notions of "fixtures" and "options." That is, when evaluating causal claims, we tend to assume certain factors in the world as being fixed, rather than variable, and this may be for pragmatic reasons. This can directly affect what parts of the system we choose to model as variables, rather than fixed parameters. Woodward does not think that this sort of relativism is a threat to his causal realism.

7.1.4. A counterfactual theory of causation. Woodward thinks of his theory as—in addition to being a manipulation theory—also a kind of *counterfactual* theory of causation. There are thus points of comparison both to earlier manipulation theories and to the counterfactual treatment by David Lewis and followers. Interventionism is a counterfactual, as opposed to an *actualist*, theory in virtue of relating the truth of causal claims to what would happen or be the case in some non-actual situation. It is furthermore an interventionist, rather than a Lewisian, counterfactual theory in virtue of the fact that the relevant counterfactual situation necessarily involves what would happen or be the case, not in a situation that is merely *different* from the actual one in some way, but one in which an intervention, specifically, occurs. (This aspect of the theory can however be somewhat complicated to understand, as we shall see below and in the next chapter.) We will look at the details of Woodward's proposal next.

7.2. Woodward's interventionist theory of causation

As stated above, Woodward's aim in *Making Things Happen* is to provide an interpretation of a qualitative notion of "direct causation"—which is denoted by the arrow in a causal diagram. He defines a causal diagram as a pair $< V, E >$ where V is a set of vertices associated with variables and E is a set of directed edges, denoting direct causation from one variable to another. Woodward focuses on deterministic causal systems. The variables V may be real-valued, categorical, binary, etc. Regardless of the types of the variables, Woodward's account is of quantitative causation, in virtue of depending on a binary notion of change or non-change, rather than the magnitudes of changes (i.e., effects). He sketches the driving manipulationist notion of the theory:

> The basic idea is that X is a direct cause of Y if and only if the influence of X on Y is not mediated by any other variables in the system of interest V in the following sense: there is a possible manipulation of X that would change the value of Y (or the probability distribution of Y) when all other variables in V are held fixed at some set of values in a way that is independent of the change in X. (Woodward 2003, p. 42-43.)

Woodward then wants to proceed by formulating necessary and sufficient conditions for something being a *contributing cause* in terms of the conditions for

direct causation (2003, p. 53). "*Total cause*" is also defined, as a special case, and it does not mean the same in Woodward's theory as what we have meant by that label so far, which was rather the minimal sufficient set of causal factors for an effect (and Woodward has no special term for this latter entity).

7.2.1. Causal relata. I noted in the previous chapter that it is common in the scientific literature to talk of causality as a relation between variables, but also that, if we think "variable" denotes a symbol in a theory, then this is an obvious shorthand for causally relating whatever it is the variables stand for in the physical world. Woodward adopts this vernacular, but says something more specific, as well, about what the causal relata of the theory are. Variables are commonly understood as denoting a quantity or property that takes on one of several mutually exclusive values for some individual—i.e., variables/values stand for determinables/determinates.

Woodward notes in particular that in the context of this theory, "causal claims tell us not that one property is associated with or necessitates another, but rather that certain *changes* in the value of a variable will produce associated changes in the value of another variable" (2003, p. 112; emphasis added). On one seemingly straightforward way to understand Woodward's notion of a change in a variable X, there exist two distinct values k_1 and k_2 in X's domain, and first $X = k_1$ for an individual and then (at a later moment) $X = k_2$ for that same individual (e.g., 2003, p. 45). Causes and effects are then *temporal change events*, or "dynamic events" in von Wright's terms (section 4.3, above). But there seems to be some ambiguity on the precise nature of these changes, which I will return to.

7.2.2. Total (net), direct, and contributing causes. While Woodward's theory takes much from the technical work by the authors mentioned in the previous chapter, an important difference is that he does not assume those main constraints on causal models that Pearl and Spirtes et al. mostly assume, namely the Causal Markov Condition and faithfulness—nor, if I understand him correctly, does Woodward assume causal sufficiency. (A model is causally sufficient if and only if any common cause of two variables in the model is also part of the model (Spirtes et al. 2000, p. 22).) Rather, the definitions of causation and of intervention are expected to do the whole theoretical work.

Woodward separates the sufficient part and the necessary part of the condition on causation suggested in the informal version of the manipulation theory, and discusses them separately (Woodward 2003, p. 45):

> **SC:** If (i) there is a possible intervention that changes the value of X such that (ii) carrying out this intervention (and no other interventions) will change the value of Y, or the probability distribution of Y, then X causes Y.
>
> **NC:** If X causes Y then (i) there is a possible intervention that changes the value of X such that (ii) if this intervention (and no other interventions) were carried out, the value of Y (or the probability of some value of Y) would change.

At this point in the presentation, Woodward thinks informally of an intervention on X as an "exogenous causal process that changes X in such a way and under conditions such that if any change occurs in Y, it occurs only in virtue of Y's [causal] relationship to X and not in any other way" (2003, p. 47). Given this understanding of "intervention," **SC** is of course "extremely plausible," as Woodward states (2003, p. 49). It looks like a logical truth.

However, in order to evaluate **SC**, and to understand if **NC** is true, for any case where an intervention is not actually taking place, we need to know when there exists a merely *possible* intervention. Woodward recognizes this fact, and I will return to the issue below, for now we will take some relevant notion of "possible" for granted.

(One more, minor point: condition **SC** is stated in a way such that the intervention *changes* Y which, if this is understood causally, is a circularity that is straightforwardly avoidable, and Woodward indeed puts the condition in a different way in other places. There he says simply that Y changes when the intervention occurs (e.g., Woodward 2014a, p. 1716).)

While **SC** is true for all cases (under the given meaning of "intervention"), Woodward argues that **NC** is not: X can be a cause of Y even though Y does not change under an intervention on X with respect to Y. This can happen when the system in question is not *faithful*, in the sense described in the previous chapter. That is, X causally influences Y along several paths, whose coefficients sum to zero. This is perhaps an *a priori* unlikely situation that could then be ignored in practice, but as it's not impossible, it makes **NC** inadequate as a general condition for the truth of causal claims. Woodward argues that a theory that aims to "cash out" the meaning of causal claims in terms of what happens under hypothetical manipulations requires a necessary condition, otherwise we "face the possibility that there is some other set of conditions, having nothing to do with facts about what would happen under manipulation of X, that are also sufficient for X to cause Y and puzzling questions about the relationship between these two sets of conditions and why they are both relevant to causation" (2003, p. 60-61). (This closely mirrors my reasons for *defining* a manipulation theory as one that implies the necessary condition, in section 1.3.) Thus, we need to find the right necessary condition.

Woodward calls X, in a situation in which X has a net non-zero influence on Y along all of its paths of influence, a *total cause* of Y.

> **TC:** X is a total cause of Y if and only if there is a possible intervention on X that will change Y or the probability distribution of Y. (Woodward 2003, p. 51.)

"Total cause" is an unfortunate choice of label for us, as I have used it to denote something like Mill's "cause"—a minimally sufficient condition for the effect. In my mind "net cause" is more descriptively salient here (attributed by Woodward to Christopher Hitchcock 2001b). I think Woodward's "total cause" derives from the common notion of "total effect"—this being the effect of one variable on another along all paths connecting them. I will mostly use Woodward's term in this chapter.

A total cause is then a special, but exceedingly common, kind of cause, for which **SC** and **NC** are the right conditions. The *general* notion of "cause," in a claim such as "X is a cause of Y," is rather that of a *contributing cause*. Woodward aims to define "contributing cause" in terms of "direct cause."

Direct causation between variables in a model is denoted by an arrow between the associated nodes in a causal diagram. Woodward argues that the notion of "direct cause" is fundamental to understanding other causal notions, and causal reasoning. For example, we need the notion to understand what will happen under combinations of interventions, and also for understanding what an intervention is (Woodward 2003, p. 52). We also need direct causation in our theory of probabilistic causation: it is required for determining what variables to control for when testing for a causal dependence, and the reliance on direct causation can also immediately be seen in the formulation of the Causal Markov Condition, when it refers to the *parents* of a variable (Woodward 2003, p. 64).

As I mentioned in the previous chapter, "direct cause" is a model relative concept, since the directness in question has nothing to do with conditions in the physical world, such as spatiotemporal contiguity—the difference between direct and indirect causation is purely a matter of what parts of the physical system have representation by variables in the model.

To determine whether X is a direct cause of Y, Woodward employs several interventions, one that changes X and others that "hold fixed" the variables along any additional, indirect causal path between X and Y. If X is then connected to Y along some indirect path that might neutralize its direct influence on Y, this counteracting indirect influence is eliminated in the test. Since it makes no difference whether we also hold fixed any variables that are *not* on a causal path between X and Y at all, the condition can be simplified: X is a direct cause of Y, relative to the variables V in a model, if and only if Y changes under an intervention on X when all other variables in V are held fixed.

Woodward recognizes that this definition of "direct cause" makes it a model relative—or variable relative—notion, but he believes this relativization to be theoretically unproblematic. He takes the reason that it is not a vicious form of relativism to be that, if X is a contributing cause of Y, it will remain a cause of Y even if new variables are introduced, that are intermediate on a path between X and Y (Woodward 2003, p. 56). This claim of *monotonicity* has been challenged by Michael Strevens, as we shall see below.

With a theory of direct causation, Woodward can proceed to define "contributing cause" in terms of it. However, while we might initially think that "contributing cause" is just the ancestral of the "direct cause" relation, Woodward states that this cannot be right, because such a definition would make causation a *transitive* relation, and he thinks that there are conclusive counterexamples to causal transitivity (Woodward 2003, p.57). He uses a well-known scenario presented by Michael McDermott in a discussion of Lewis's counterfactual theory of causation (McDermott 1995). In the imagined situation, someone presses a button to detonate a bomb. However, before that

happens a dog bites the hand of the bomber, leading her to press the detonation button with her left hand instead of the right. The dog bite thus causes the bomber to press the button with her left hand, and the bomber pressing the button with her left hand causes the bomb to explode. But since the bomber pressing the button with her right hand would also have brought about the same effect of the bomb detonating, whether the dog bites the bomber or not makes no difference to this effect. Thus, under a counterfactual theory, the dog bite is not a cause of the bomb exploding, and transitivity is defeated.

We could model the situation as in figure 7.2.1, where B is a variable indicating whether the dog bites the bomber's right hand ($B = 1$) or not ($B = 0$), E indicates whether the bomb explodes ($E = 1$) or not ($E = 0$), and P models the button-pressing and takes one of *three* values, depending on if she presses the button with her right hand ($P = 1$), her left hand ($P = 2$), or not at all ($P = 0$). I have added a binary variable T for clarity, which stands for whatever event ($T = 1$) causes the bomber to perform this act of terror. The equations of the model are seen in table 7.2.1. We see there that there is *some* value of T, namely 1, such that if we fix T to that value, then P will change when we change B by an intervention. But there is *no* value of T such that if we fix T to that value, E will change when we change B by an intervention. Clearly there is some change in the value of P (namely any change from 0, to 1 or 2) that, if brought about by an intervention, will change the value of E. Thus, B is a cause of P and P is a cause of E, but B is not a cause of E.

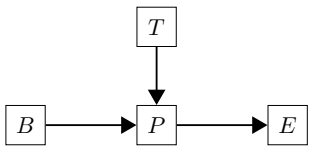

FIGURE 7.2.1. Diagram for "Dog Bites Bomber" (counterexample to causal transitivity).

The non-transitivity of this case depends formally on that E is not causally sensitive to every possible change in P. That not *any* change in a cause variable is followed by a change in the effect variable is something Woodward generally embraces in the theory. The condition is that there is *some* intervention, that sets the cause variable to *some* value, such that the outcome changes under that intervention. In the model of Dog Bites Bomber, P is fined-grained enough to preserve information about whether the dog bites or not, but this information is not preserved in E. There is thus a more course-grained variable P' that we could have used instead, that indicates only whether the bomber presses the button or not. T would be a cause of this variable, and E would be an effect of it, but B would not be a cause of it.

1. $P \stackrel{c}{=}$ 0 *if* $T = 0$, 1 *if* $T = 1 \wedge B = 0$, 2 *if* $T = 1 \wedge B = 1$.

2. $E \stackrel{c}{=}$ 1 *iff* $P = 1 \vee P = 2$.

TABLE 7.2.1. Equations modeling the Dog Bites Bomber case (figure 7.2.1).

It may be interesting to note that since non-transitivity along a causal path from X to Y to Z depends on there being a range of values in the intermediate cause Y such that Y is sensitive in this range for some changes in X, and Z is *not* sensitive to changes in Y within that range, it follows that if all causal dependencies in the system are continuous functions between real-valued variables—as in the common linear cases as well as those giving probabilities of binary outcomes—causation *is* transitive. Thus, to the extent that these are the common cases in the scientific context, causal transitivity would tend to be the norm there.

Accepting that causation is not in general a transitive relation, Woodward can't define "contributing cause" as the ancestral of "direct cause." Rather, he adds the condition that if X is a contributing cause of Y along some path P, then Y changes under some intervention on X when every variable in V that is not on P is held fixed at some value. We can now state Woodward's theory, which he calls **M** for "manipulability theory" (Woodward 2003, p. 59; emphasis and some whitespace added):

> **M:** A necessary and sufficient condition for X to be a (type-level) *direct cause* of Y with respect to a variable set V is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V.
>
> A necessary and sufficient condition for X to be a (type-level) *contributing cause* of Y with respect to variable set V is that
> (i) there be a directed path from X to Y such that each link in this path is a direct causal relationship; that is, a set of variables $Z_1 \ldots Z_n$ 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 $\ldots Z_n$, which is a direct cause of Y, and that
> (ii) there be some intervention on X that will change Y when all other variables in V that are not on this path are fixed at some value. If there is only one path P from X to Y or if the only alternative path from X to Y besides P contains no intermediate variables (i.e., is direct), then X is a contributing cause of Y as long as there is some intervention on X that will change the value of Y, for some values of the other variables in V.

In his exchange with Michael Strevens, Woodwards clarifies or modifies the condition for contributing causation. He makes the notion of contributing causation *not* relative to the set of variables V in a certain causal model, by existentially quantifying over variable sets. Specifically, he takes **M** to state the conditions under which a variable X is *correctly represented* as a contributing cause within a certain model. He then adds that "X is a contributing cause of Y *simpliciter* (in a sense that isn't relativized to any particular variable set V) as long as it is true that there exists a variable set V such that X is correctly represented as a contributing cause of Y with respect to V" (Woodward 2008, p. 209). We may thus add the following definition of "contributing cause simpliciter":

CCS: X is a contributing cause of Y simpliciter *if and only if* there exists a variable set V such that X is a contributing cause of Y with respect to V.

This is a theory of *general* causation. That is to say, it causally relates type-level things, not particular events. Woodward presents a theory also of causation between particular events and, like Pearl, he calls this the theory of "actual causation." In this account, Woodward follows Christopher Hitchcock's and Judea Pearl's treatments (Woodward 2003, note 41 to ch. 2). He also states in an interchange with Michael Strevens that actual causation "occupies a peripheral role" in *Making Things Happen* (Woodward 2008, p. 197). For these reasons, and for reasons of scope, I will not discuss Woodward's theory of actual causation. Suffice it to say that the main difference between actual and type-level causation is as follows. While the question whether X is a type-level cause of Y in some system hinges on whether there are *some* values of certain other variables in the system such that Y would change under an intervention on X when those other variables are held fixed at those values, the question whether an actual X-event (a change in X on a particular occasion, in a particular instance of the system) was a cause of an actual Y-event hinges on whether Y would change under an intervention on X while those other variables in the system are fixed, not at some, but at their *actual* values. But, as with the definition of "contributing cause," there are complications, in this case coming in the form *redundant causes*, which affect the complexity of the definition. For more, see Woodward 2003, sect. 2.7, and Strevens 2007, Woodward 2008, Strevens 2008.

7.2.3. Interventions. So far we have relied in our definitions on Woodward's informal characterization of an intervention on X (that tests whether X is a cause of Y), which stated that an intervention is an "exogenous causal process that changes X in such a way and under conditions such that if any change occurs in Y, it occurs only in virtue of Y's [causal] relationship to X and not in any other way" (2003, p. 47). Woodward also thinks of interventions as "an ideal experiment designed to determine whether X causes Y," echoing the econometrician Guy Orcutt from the last chapter (Woodward 2003, p. 46).

The connection is plain, then, between interventions and actual manipulations in successful scientific experiments. Such manipulations are taken to be interventions. But it is a feature—perhaps the preeminent feature—of Woodward's theory that interventions are understood in a perfectly naturalistic way, such that free actions by humans or other agents have no special theoretical role, except as and when they satisfy the conditions for being an intervention. This eliminates the anthropocentric implications of earlier manipulation theories of causation, and more clearly accommodates the causal realism Woodward advocates. It does however require that interventions are ultimately given a definition in causal terms, and thus any hope of a reductive analysis of causation must be abandoned.

Guided by the informal characterization, as well as the goal of theoretically identifying an ideal experimental manipulation, Woodward introduces four conditions on an event being an intervention. He begins by defining an

intervention variable in the context of a causal model, and goes on to define an actual *intervention event* in terms of this. I will begin by explaining the conditions on an intervention variable, before stating its definition.

An intervention modeled by a variable I is meant to test whether some variable X is a cause of some variable Y, and the intervention is therefore relativized to these variables, and we say that I is an intervention on X with respect to Y, if it satisfies the following conditions. *Firstly*, I must clearly be a cause of X. *Secondly*, I must be the *only* cause of X. This follows the rule in Pearl's treatment for modeling an intervention, where we were instructed to delete all other arrows in the diagram that point at X. But Woodward expresses himself directly in terms of the causes of X, so that his condition is not relative to some particular causal diagram. He calls an intervention on X a "switch," meaning that for *some* values of the intervention variable I, all the causal influences on X in the unmanipulated system are "switched off" in the manipulated system. (This condition is relaxed in a later amendment that I will introduce below.) *Thirdly*, the intervention may not affect the outcome Y independently of X. This can clearly happen, when a manipulation has unintended consequences that independently affect the outcome of an experiment. A violation of this condition corresponds to there being a path from I to Y that does not pass through X, in the model of the manipulated system. *Finally*, I may not be correlated with any causes of Y in the system, that are not either causes of, or caused by, X. If this were the case, it could explain how X and Y correlate under the manipulation of X, even though X is not a cause of Y. In terms of a model, the condition requires that, if Ycorrelates with X under this intervention, all other variables that I correlates with are on some path from I to Y that goes through X. Here now are the conditions on an intervention variable, as Woodward states them (2003, p. 98). "Cause" in the conditions means "contributing cause."

IV: I is an intervention variable for X with respect to Y *if and only if*

I1. I causes X.

I2. 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.

I3. Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the I-X-Y connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.

> *I4*. I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X.

In consideration of work mainly by Frederick Eberhardt and Richard Scheines (2007), Woodward has later weakened the conditions. It has been shown that an effect of X on Y can be identified even if an intervention does not eliminate other causal influences on X within the system, but merely changes its probability distribution while leaving the existing causes of X intact. Eberhardt and Scheines call this a "parametric" intervention, or a "soft" intervention, the latter seeming to have become the more popular term (Woodward 2015a, 2015b, 2017). Accordingly, condition *I2* can be abandoned when it is soft interventions that are assumed.

Woodward points out that if we assume the Common Cause Principle (which is implied by the Causal Markov Condition), then condition *I4* can be reduced to the requirement that I does not have a common cause with Y. This is the case since under the Common Cause Principle, a correlation between I and some cause Z of Y in the system must be due either to that I is a cause of Z or that Z is a common cause of I and Y. In the first case, Z is either on a causal path between I and Y that passes through X, in which case all is well, or Z is not on a path that includes X, but this situation in turn violates condition *I3*. That leaves only the possibility that Z is a common cause of I and Y.

We may note two differences here, to Pearl's treatment of interventions, that I think are connected. The first is that Woodward relativizes the intervention to Y. The second is that Woodward's exogeneity condition is weaker than Pearl's. The first difference, I think, explains the latter. By relativizing the intervention to Y, Woodward can require (*I4*) that the intervention is independent only of any other causes of Y, while Pearl requires that the intervention is exogenous to the causal system as a whole. (Assuming that the system is causally sufficient–i.e., if two variables in the system have a common cause, then that cause is in the system—Pearl's exogeneity condition implies that the intervention has no common cause with the purported effect.) We have thus encountered three different scopes and strengths of exogeneity in the theories we have reviewed. In order of decreasing scope and strength: manipulations under the Ramseyan assumption, as described in subsection 5.6.3 above, are what we may call "globally exogenous"—they have no causes at all beyond the agent herself or her intention to act; Pearl's interventions lack causes in the modeled system—we might call this "system-relative exogeneity"; Woodward's interventions, in turn, have no causes among the set of causes of the outcome Y—which we might call "outcome-relative exogeneity." Woodward's exogeneity condition is thus of the weakest type, and employs the most specific causal information about the system in its formulation.

Woodward goes on to define an actual intervention event (2003, p. 98).

> **IN:** 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.

A literal reading of **IN** makes every possible intervention actual, but this isn't Woodward's meaning, and I will return to this issue in the next chapter. For now, we may think of an intervention event as being a type-level event cause of the event of X taking on some value.

As Woodward's conditions for an intervention refer to (contributing) causes, and his theory **M** of causation relies on interventions, there is no way to reduce a claim that X is a cause of Y to a claim about non-causal facts. Woodward acknowledges this, and insists that "a manipulability theory can provide a non-trivial *constraint* on what it is for a relationship to be causal without providing a reductive analysis of causality" (2003, p. 106).

However, Woodward also argues that the theory is not viciously circular, because the conditions under which X is a cause of Y don't involve specifically X being, or not being, a cause of Y. They involve only causal relations between other pairs of variables (Woodward 2003, p. 104). In *some* sense there is no circularity in the theory, then, but whether this is a satisfactory situation seems contingent on what we take it to be a theory about. Even if the theory provides non-circular conditions for X being a cause of Y specifically—but conditions that rely on facts about other causal relations—have we really succeeded in "cashing out" the content of causal claims in terms of interventions (as Woodward aims to do (2003, p. 60-61)? The circularity issue is explored further in Michael Baumgartner's criticism, which we will review below.

7.2.4. Modularity and the possibility of intervention. In this section I will tie together several discussions by Woodward that I believe are closely interconnected. Mainly they deal with the possibility of intervention, and with the modularity of causal systems.

First of all, let's note a difference between Woodward's (and Pearl's) treatment and that of the earlier manipulationist theories, when it comes to causes that cannot be manipulated in practice. In section 1.3 I proposed four different ways of understanding the general manipulationist condition in such cases. Woodward's (and Pearl's) theory goes for option three in this list: it insists that the cause *is* manipulable—or intervenable—in the theoretically relevant sense. Gasking, von Wright, as well as Menzies and Price opted rather for the forth option: the cause is at least connected to manipulable things by a special relation. E.g., the cause is of the same sort as some manipulable thing, or it is composed of manipulable parts, or it shares some intrinsically non-causal properties with manipulable things. The manipulable thing(s), that the unmanipulable cause has the relevant relation to, in turn satisfies the manipulationist condition on causes. (Collingwood, interestingly, also insists, like Woodward, that all causes are manipulable—but that they are *practically* manipulable. He simply rejects practically unmanipulable causes all-around.)

The manipulationist necessary condition in Woodward's theory thus implies that if X is a cause of Y in a system, then there is a possible intervention on X with respect to Y in that system. As I mentioned above, in order to evaluate this condition, we need to know what "possible" means in this context. Woodward recognizes this requirement: he says that the theory "needs to make clear how such counterfactuals are to be understood and what their truth

conditions are. In particular, we need to know just what sort of possibility we should be envisioning" (2003, note 11 to ch. 2). We can exclude immediately that it is a matter of *practical* possibility since Woodward, like most of his manipulationist predecessors, wants to allow for things like the presences in certain locations of planet-sized objects being causes. Plausibly, the next degree of decreasing modal strength is physical, or nomic, possibility. Woodward thinks that this, also, is too strong.

By a physically possible event we here mean an event that is compatible with the physical laws that govern the system, under some initial conditions that are themselves allowed by the laws. Woodward produces two examples to argue that the possible intervention cannot generally be expected to be physically possible. In these examples Woodward emphasizes that the possible intervention is here a physical *change event* or process, that satisfies the conditions for being an intervention.

One example is that of a cause `C` that is a truly random event and that could therefore not be caused to occur by anything (Woodward 2003, p. 130). Since `C` cannot be caused, condition *I1* on an intervention on `C` is unsatisfiable. Woodward takes this to be at least a coherent idea, but I think we can even produce a real-world example. There are true random number generators, that produce a number based on the nuclear decay of some radioactive source such as americium-241. In such a device, emissions of alpha particles are truly random events that can be detected and used to produce a number that is theoretically impossible to predict given any information about the preceding states of the system. Thus, that a physical process may set the alpha emission in the device to some particular value is excluded by the laws of physics as we know them. Still, we would probably want to say that the alpha emission causes what number shows up on the computer screen. (Another familiar fictional example is the mechanism that kills—or not—Schrödinger's cat.)

In a different example, we assume it to be true that

> [c]hanges in the position of the moon with respect to the earth and corresponding changes in the gravitational attraction exerted by the moon on various points on the earth's surface causes changes in the motion of the tides. (Woodward 2003, p. 129.)

The problem now is not condition *I1*, but condition *I3*: that an intervention does not affect the outcome except by way of the presumed cause. If we restrict ourselves to physically possible events, then an event that moves the moon may necessarily involve forces of such magnitudes that they will affect other parts of the system besides the moon's location, including the motion of the tides on Earth. That this *may* be true is enough, Woodward thinks, to prevent us from assuming that there is always a physically possible event that is "surgical" enough to affect only the intended target of the intervention.

Although Woodward doesn't describe it in such terms, the moon/tide example—under the assumption that a surgical intervention on the position of the moon with respect to the motion of the tides is physically impossible—looks like a case of *failure of modularity*, when "modularity" is understood in a

certain way. We encountered the modularity notion in the last chapter (section 6.3). Woodward defines modularity as follows.

> [A] system of equations will be modular if it is possible to disrupt or replace (the relationships represented by) any one of the equations in the system by means of an intervention on (the magnitude corresponding to) the dependent variable in that equation, without disrupting any of the other equations. (Woodward 2003, p. 48.)

That causal systems are generally modular with respect to causes and their effects is required by condition *I3* on interventions, and thus implied by the theory. Moreover, the reference to possibility in Woodward's characterization makes it clear that modularity is a modal, or counterfactual, notion. Due to how the possibility of intervention depends on modularity, the sense of "possible" in the definition of modularity can at least not be *stronger* than that which occurs in the conditions for intervention. The moon/tide example is a failure of modularity if "possible" in this definition is taken to mean "physically possible." On this interpretation of "possible," the moon/tide system is not modular if it's not physically possible to cause a change only in the position of the moon, and not at the same time in the tides, independently of the effect on the moon.

Woodward suggests as a motivation for accepting the modularity of causal systems that we may take it as a conceptual truth of causal mechanisms that they are independent in the way stated in the definition of modularity (2003, p. 48). Thus, any causal model in which it is not possible to intervene on a cause X with respect to one of its effects Y, without doing violence to the dependencies modeled by equations other than X's, simply hasn't gotten the causal mechanisms in the system right. The moon/tide example, if correct, then shows that "modularity" cannot refer strictly to a physical possibility.

Returning to Woodward's discussion about the possibility of intervention, he suggests the following substitute for thinking of interventions as physically possible change events.

> [A]s long as there is *some basis* for assessing the truth of counterfactual claims concerning what would happen if various interventions were to occur, it doesn't matter that it may not be physically possible for those interventions to occur. (Woodward 2003, p. 130; my emphasis.)

This basis consists in at least two things: a "coherent conception" of what it is to change the variable intervened on; and "some grounds for saying what the effect, if any, on [the outcome] E would be of changing just [the purported cause] C and nothing else" (Woodward 2003, p. 130-131). In the moon/tide example, Woodward thinks that Newtonian theory itself provides the latter. This is reminiscent of what Pearl said in response to Cartwright's claim that not all causal systems are modular. He stated there that what is required is literally a *symbolic* intervention.

> Symbolic modularity does not assume physical modularity. [...] Symbolically, one can surely change one equation without altering others and proceed to define quantities that rest on such "atomic" changes. (Pearl 2009, p. 364-365.)

Based on Pearl's characterization, we might say that what is required for an intervention on a cause to be possible is that we have a causal theory, of the kind we have encountered here, in which the cause is modeled as a variable. As I understand this account, if the theory is also true, then the implied intervention counterfactual is true, and this is how these counterfactuals are interpreted. But this makes the connection between an intervention viewed as a *physical change event*, and the possible intervention implied by the theory, obscure. The latter must perhaps then rather be understood as a possible *symbolic operation* on a certain correct representation of the system in question.

That Woodward does not think of interventions in quite these terms is suggested by his conclusion that

> an intervention on X with respect to Y will be "possible" as long as it is logically or conceptually possible for a process meeting the conditions for an intervention on X with respect to Y to occur. (Woodward 2003, p. 132.)

This seems to retain the imagery of a possible physical event, rather than a possible symbolic operation on a theory. (In more recent writings, too, Woodward has proposed that the right sense of "possible" in his conditions is at least logical or conceptual possibility (2015b, p. 3583).)

What I think we can conclude from this is that the requirement that there is a possible intervention on any cause in any system, that is given by the conditions in **IV**, is at least a weaker constraint on causes than we might at first have imagined. If we look again at what Woodward says about the methodological utility of the interventionist account, the stronger constraint appears to be rather the first one alluded to above: that we have a coherent idea of what a *change* in the relevant variable would be. (I will return to this issue in the next chapter.)

7.3. Critique: circularity, relativity, and realism

In this section I will focus on three issues that I find central to the question of how we are to understand Woodward's proposal—that is, to the question of what Woodward's theory is a theory *about*, more precisely. These are, firstly, the circularity in the definitions of the theory. Woodward has argued that there is at least no *vicious* circularity in his definitions. This may give us some clues as to Woodward's intentions, especially in light of some arguments to the effect that the circularity *is* fatal under some ways of understanding the theory. Secondly, there is the issue of model relativity, which reasonably indicates something about what it is that makes the conditions of the theory true, in cases when they are. Thirdly, there is the general question of causal realism, and how Woodward's theory appears to fair under such an expectation.

7.3.1. Circularity and epistemic or semantic grounds. To briefly review the potential circularity issue, Woodward defines "X is a direct cause of Y" in terms of a possible intervention on X with respect to Y, and what would happen if this intervention was carried out. "Intervention" is defined in explicitly causal terms, in turn. Woodward provides essentially two points of defense with respect to the circular nature of his theory: firstly, that the theory can be informative without being reductive and, secondly, that the conditions under which X causes Y do not themselves directly depend on whether X causes Y. These points have been generally conceded among critics. But several writers have nevertheless perceived the circularity as problematic.

Clark Glymour has said that the definition of "direct cause" in the theory is "ill-founded, not circular: it could never be applied to determine direct causes ab initio" (Glymour 2004, p. 785). (I will return to the relation between circularity and ill-foundedness in the next chapter, section 8.4.3.) Henk W. de Regt in his review considered the circularity a problem for Woodward's goal of providing an interpretation of causal claims. The fact that whether X is a cause of Y is not directly involved in determining whether X is a cause of Y, while other causal relations are involved, he considered an acceptable defense only if the theory "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).

Especially in later writings Woodward has emphasized the pragmatic, or "functional" (also "methodological") side of his theory. Some critics have claimed that the circularity is a problem for the theory viewed in this way, as well. Michael Strevens has argued that, if we attempt to complete the process of determining the causal relationship between two variables X and Y in accordance with Woodward's definitions, then we will either run into a "dead end" or we will find that the result depends on whether X is a cause of Y after all (Strevens 2007). The most elaborate analysis of Woodward's theory along these lines, and from an explicitly methodological point of view, is by Michael Baumgartner (Baumgartner 2009). Baumgartner aims to show that the circularity in the definitions lead to a vicious infinite regress in the application of the theory.

Baumgartner starts from the assumption that Woodward's theory is to provide us with a practically feasible *method* for determining whether X is a cause of Y in some causal system. He then observes that for this to be the case, it must be possible for us to identify a variable I_1 as an intervention on X with respect to Y, in that system. This means in particular verifying that the conditions **IV** are satisfied by I_1 relative to X and Y. Condition *I1*, then, requires that I_1 is a cause of X. Knowing this in turn requires that we identify an intervention I_2 on I_1 with respect to X. This begins an infinite regress. Condition *I3* proscribes that I_1 is not a cause of Y along some path not containing X. To know whether I_1 is a cause of some variable that is a cause of Y and that is not connected to X, we must identify an intervention I_3 on X that holds that variable fixed. This initiates the same regress as before. Hence, we can see that it is not possible to identify an event as an intervention "*ab initio*," within the theory. It's easy to imagine that completing this procedure,

for each of the conditions, will generate more regresses of this type, but these two suffice for demonstrating that the theory does not on its own provide us with a feasible method for discovering causes (Baumgartner 2009, p.181).

Baumgartner proceeds to consider two possible ways of avoiding these "identification regresses." We might rely on prior causal knowledge about something being an intervention on X with respect to Y. Or, we might just be assuming that something is such an intervention. He goes on to show that, given certain common assumptions, neither of these strategies can be successful.

As to employing prior causal knowledge, it does indeed seem obvious that we sometimes know that an event is an intervention of the right type, for example when we use coin tosses or some other intrinsically random process to randomize a trial. Baumgartner suggests that this prior knowledge may in turn have two kinds of justification in the present theory. Either it is justified by direct application of the conditions in **M** and **IV**, or it is justified rather by some available heuristic. In the first case, there is an *epistemic* regress that mirrors the identification regress we have just described: to *know* that I_1 is an intervention on X with respect to Y requires that we know that there is an intervention I_2 on I_1 with respect to X (condition *I1*), and so on. If we think our prior knowledge is justified instead by some suitable heuristic, Baumgartner argues that this in turn cannot be justified unless, for at least *some* cases, when the heuristic indicates that an event is an intervention, we can confirm that it *is* an intervention. But the identification regress prevents us from ever doing this, for any case whatever. Baumgartner thinks that these *justification regresses*, together with the belief that we *do* know that some events are interventions, implies that we as a matter of fact do not understand causation in the way Woodward's theory proposes (Baumgartner 2009, p. 187).

Towards the end of this monograph, I will sketch an account of causal knowledge that I take to address Baumgartner's point here, in a general sense. This account nevertheless makes essential use of some features of interventionism.

The second way of ending the identification regress was to *assume* that an event is an intervention. Such an assumption can indeed be employed for the purpose of showing that a certain causal inference on a model is *valid*, and a lot of work in causal modeling theory is precisely of this nature (e.g., Spirtes et al. 2000, Pearl 2009). Baumgartner shows that for some distribution of variable values in an observed system, that is compatible with several distinct causal models (as is always the case), different assumptions about what variables are interventions relative to what other variables lead to different conclusions about which one of the models truly represents the system (Baumgartner 2009, p. 190-191). This result is uncontroversial, but it means that if the theory is to provide a method for acquiring causal knowledge about the world, it must be possible to justify our choice of causal assumptions, and Baumgartner has already argued that this cannot be done.

Thus, the implication is that, even if Woodward's theory does impose some constraints on the relationship between intervention and causation, it cannot

provide a method that is sufficient for acquiring causal knowledge. Baumgartner suggests, moreover, that there are other philosophical accounts of the causal relation that do allow us to identify a variable as satisfying the conditions in **IV**, in a finite number of steps, such as Suppes's probabilistic theory or, at least in certain cases, a mark transfer or process account in the style of Salmon and Dowe.

Baumgartner's argument depends on the regresses of identification and justification never terminating or—viewed from the other direction—Woodward's theory not providing a base from which we can build up our causal knowledge in accordance with the interventionist criteria. Victor Gijsbers and Leon de Bruin have recently proposed that this base can be supplied by a primitive agency theory of causation (Gijsbers and de Bruin 2014). Their proposal has two major parts. One is that interventionism is as a matter of fact a sophisticated generalization from a primitive agency theory of causation. The other is that the primitive agency theory can generate a base of causal knowledge that allows interventionist causal inferences to get off the ground, and that also itself conforms to the more developed interventionist conditions for causation, once these are theoretically available. By proposing that the generalization of "cause" to unmanipulable causes is the interventionist theory, this account differs importantly from the earlier manipulationist theories, although it also has marked similarities to all of them. Perhaps it has an especially close relation to Menzies's and Price's claim that their theory isn't circular, because we have, as it were, direct acquaintance with the notion "to bring about," which they then use in their analysis of "to cause" (see section 5.4.2). The account of causal knowledge that I sketch toward the end of the final chapter agrees with the second part of Gijsbers's and de Bruin's proposal, but doesn't imply or involve a generalization of a "primitive agency theory of causation."

7.3.2. Variable relativism, monotonicity, and causal realism. One criticism of Woodward's theory, by Michael Baumgartner, highlights the difficulty in identifying something as an intervention, that is due to the circularity in the definitions of causation and intervention. The objection has close similarities to one made earlier by Michael Strevens (Strevens 2007). In that same paper, Strevens also problematizes the fact that, in Woodward's theory, causal relations appear to be relativized to the choice of variables in a causal model. If, according to the theory, what causal relations there are depends on what one has chosen to model as a variable, or is capable of modeling, then this may undermine Woodward's claims of causal realism. The discussion leads Woodward to reformulate his condition on contributing causation (in a way I have already described, in condition **CCS** above). Strevens illustrates how he understands the variable relativism problem with a straightforward example:

> Assume, for example, that the amount of expensive bottled water you drink and your chances of succumbing to heart disease are correlated, because they share a common cause, say, the consumption of salty food. Consider a causal network containing variables representing water-drinking and heart disease but not salt consumption. Because Woodward's definition of

> an intervention is implicitly relativized to the variables in a network, increasing the amount of bottled water you drink by increasing your consumption of salty foods will count as an intervention relative to the salt-free network (due to the invisibility, within the network, of the "side effects" of the salty strategy for drinking more. Thus, because "intervening" on your water consumption in this particular way will increase your chances of getting heart disease, water consumption will count as a cause of heart disease [...] (Strevens 2007, p. 243.)

It seems to me that this example fails to show Strevens's point, for the following reason. (A similar objection is made in McCain 2015.) If the intervention on water drinking is salty food eating, then it is salty food eating that is represented by I in the diagram of the system under intervention. It is I that must satisfy the conditions in IV. And under the assumption that salty food eating is a cause of heart disease independently of water drinking, then this I fails to be an intervention on water drinking with respect to heart disease (by violating condition *I3*). This would be determined by intervening on I while holding water drinking fixed, and as the example is presented, this would change the probability of heart disease. To be precise, there is no model of the system under this particular way of intervening on water drinking, that does not include a variable for salty food eating—as it is the very intervention in question. Thus, even if variable relativity is a problem, this example can't show it.

In his response to Strevens, Woodward explains that he never intended "contributing cause" or "total cause" to be relativized notions. He then supplies an interpretation or correction of his definition of "contributing cause." He states that X is a contributing cause (simpliciter) of Y if and only if *there exists* a set V of variables such that X is a contributing cause of Y relative to V (Woodward 2008, p. 209). This is condition **CCS** in the previous section. Moreover, he states that his theory is such that once a model has sufficient variables for identifying some contributing causal relation in it, no further addition of variables to the model can make that relation go away:

> Within a directed graph representation, arrows between variables can disappear as we add new variables [e.g., the arrow from X to Y disappears when we add an intermediate factor to the model], but a parallel claim is not true of the representation of contributing and total causal relationships. (Woodward 2003, p. 209.)

This amounts to a certain kind of monotonicity in the relativized causal relation. In his follow-up, Strevens recognizes that this definition of "contributing cause," together with the claim that relativized contributing causation is monotonic, indeed eliminates the variable relativity of contributing causation simpliciter. He notes that under these conditions, the definition of "contributing cause simpliciter" is equivalent to one that says that X is a contributing cause simpliciter of Y if and only if it is a contributing cause of Y relative to a model that contains *all* of the variables of the system. (But he also notes

that "all variables" is a problematic notion within this framework, especially if physical causal processes are continuous.) This, again, looks just like what we would want in a realist theory of causation. But Strevens now thinks that it is rather the monotonicity claim that is suspect. The question is whether the introduction of new causal relations in an amended model, as new variables are added, can in itself make old relativized causal relations disappear. If this is the case, then monotonicity is violated. And if monotonicity is violated, then the variable relativism in Woodward's theory might again constitute a problem for his realism. Strevens presents an argument to the effect that monotonicity does not hold. As this potential issue looks to be of considerable importance to Woodward's account, and in particular its realist interpretation, I will recount Strevens's argument in some detail.

He gives the following overview of the argument (Strevens 2008, p. 175-176).

> (1) Adding variables to a variable set can sometimes make relativized causal relations appear (as monotonicity allows).
> (2) A variable's counting as an intervener depends on the *non-existence* of certain relations of relativized causation.
> (3) Thus (from (1) and (2)), variables may lose their status as interveners as other variables are added to the variable set.
> (4) A variable's status as a relativized cause requires the existence of an intervener with respect to which a certain further condition is satisfied. If a variable loses its status as an intervener, then, other variables may lose their status as relativized causes.
> (5) Thus (from (3) and (4)), variables may lose their status as relativized causes as other variables are added to the variable set.

We will examine the steps of the argument below. By way of illustration, Strevens also provides an example of a system that he argues is such that when its model is extended with new variables, a relativized causal relation disappears in the new model.

The example is an extension of Strevens's original one, quoted above. In this case, eating salty food (S) causes a person to either drink red wine (W) or bottled water (B), but not both. Moreover, eating salty food also causes a hardening or the arteries (A). Drinking bottled water has no effect on the chance of heart disease (H), but hardening of the arteries has a positive effect on H and drinking red wine as a negative effect on H. Finally, it happens to be the case that the relative frequency of W conditional on S is such W exactly cancels out the effect of A on H. The first question now is whether S qualifies as an intervention on B with respect to H, in a model that only includes the variables $V = \{S, B, H\}$. Strevens argues that it does.

(1) S is a cause of B (*I1*).
(2) There are no other causes affecting B (*I2*).

(3) S is not a cause of H independently of B, relative to V (*I3*).
(4) S does not correlate with any other causes of H (*I4*).

The second question is now whether the probability of H changes under this intervention S on B. Strevens argues that it will, in the following way. He takes it that when S is employed as an intervention on B, only cases in which B occurs will be counted. Cases in which B does not occur (and, unknown to the experimenter, W occurs instead), will be disregarded as failed interventions. But in the set of cases in which B occurs, W does not occur, so that in this set the effect of A on H is not eliminated, and thus the probability of H changes under the intervention S on B. Thus, B satisfies the conditions for being a cause of H in this model of the system. The final question is whether the causal relation from B to H can go away as we add variables representing further factors in the system.

Strevens notes that, if X is causally connected to Y along two paths that cancel out, and the model does not contain any intermediate variables along those paths, then Y will not change under an intervention on X with respect to Y while all other variables are held fixed, and thus X will not count as a cause of Y. Once at least one variable intermediate on a path from X to Y is introduced, we must hold it fixed when testing for a direct causal relation, and that will neutralize one part of the counteracting influences on Y, so that Y indeed changes under the intervention, and X then qualifies as a relativized cause of Y.

Thus, if the model of Strevens's example is supplemented with either A or W, S can be seen to be a relativized cause of H independently of B, and it no longer qualifies as an intervention on B with respect to H. As there are by assumption no other events that would qualify, there no longer exists an intervention on B with respect to H such that H changes under that intervention, and B is consequently not a cause of H. Monotonicity is defeated.

Kevin McCain has recently proposed a refutation of Strevens's argument (McCain 2015). The objection is essentially the same as the objection to Strevens's first example. There, the problem was that salty food eating (S) does not qualify as an intervention on water drinking (B) with respect to heart disease (H) in the very first model ($V = \{S, B, H\}$), so B doesn't qualify as a relativized cause of H there either. The reason is that, in that first example, if S is varied while B is held fixed, H varies. S is thus a direct cause of H, violating *I3*. Strevens's second example ultimately has the same problem, although it takes a slightly more complicated form. The issue is thus point (3) in the above justification of S being an intervention on B with respect to H relative to $\{S, B, H\}$. The problem for Strevens's argument is that the A-path and the W-path between S and H only cancel each other out when we *leave B alone*. Strevens's argument depends on this being the case. But it means that when we test whether S is a relativized cause of H independently of B—and thus change S while we hold B fixed—the probability of H will change. Specifically, if we hold fixed $B = 0$, then salty food eaters will *only* drink red wine, and this will result in a net negative effect on the probability of H, and if we hold fixed $B = 1$ then salty food eaters will *never* drink red wine, and this will result in a

net positive effect on the probability of H. Thus, S is not an intervention on B with respect to H in a model with variables $\{S, B, H\}$, and B is, consequently, not a cause of H there.

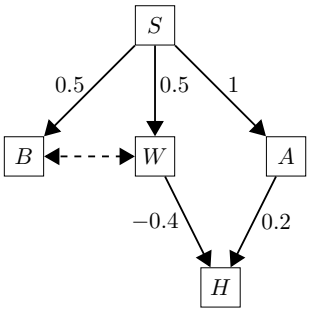

FIGURE 7.3.1. A (non-standard) model of the second example in Strevens's argument against monotonicity in Woodward's relativized causal relation.

While this seems sufficient for refuting Strevens's argument, I think we can add some interesting details to McCain's analysis. Figure 7.3.1 shows a causal model compatible with Strevens's description. (In particular, the products of the path coefficients for the A-path and the W-path sum to zero.) Firstly, the system described by Strevens does not satisfy the Causal Markov Condition (CMC). The reason is that at the step from S to either B or W (and not to both), this choice is indeterministic, meaning that B and W are non-causally dependent on each other. This dependence is not eliminated by conditioning on their common cause, and only parent, S, which violates CMC. This dependence is indicated in figure 7.3.1 by the non-standard dashed double-headed arrow between B and W. This does not on its own refute Strevens's argument, since Woodward nowhere in his theory assumes CMC. But, as a consequence, in one straightforward way of modeling this system, any event that holds B fixed will change the *coefficient* of the path from S to W. (When we fix $B = 1$, for example, the probability of W given S is zero.) Woodward has explicitly stated that he takes an event that fixes the value of X, and changes an outcome Y by changing a coefficient in some equation other than X's, as violating condition *I3* (Woodward 2003, p. 99). Strevens's example thus fails at this point, too, at least under this way of modeling it (and I don't know of any other).

Returning to Strevens's outline of his argument, (1) and (2) seem incontrovertible. But (3) does not follow from them, since it may be that the causal relations that appear in more variable-rich models are never those that would disqualify an event from being an intervention, that previously did so qualify. Strevens's has moreover failed to provide such an example. Thus, at least one attack on this kind of monotonicity of the causal relation in Woodward's models seems to have been averted, and I think that this is as important a result as monotonicity is to Woodward's causal realism.

Next, in his review of *Making Things Happen*, Peter Menzies presented two sources of concern with respect to whether Woodward's theory is as realist as he wants it to be. Firstly, Menzies wonders how Woodward would answer "the Euthyphro question: Does an intervention on X change Y because X causes Y, or does X cause Y because an intervention on X changes Y?" (Menzies 2006, p. 824). This speaks directly to the question of grounds, and Woodward's position in *Making Things Happen* appears ambiguous. Menzies observes that

Woodward sometimes speaks as though he wants to answer in the former way, as when he states that "'When an intervention changes C and in this way changes E, this exploits an independently existing causal link between C and E" (Woodward 2003, p. 132). However, there is apparent tension between these occasions and Woodward's overall aim of producing a theory of the content of causal claims that depends on truths about possible interventions—as well as his general sympathy for the manipulationist approach to causation. At this point, I think it is clear that Woodward specifically does not want to answer the Euthyphro question—and if there is a problem here, with respect to that question, it is just that.

Secondly, Menzies thinks that the variable relativity of the definitions in Woodward's theory "raises some serious questions about the extent to which a full-blooded realism about causation can be sustained" (Menzies 2006, p. 825). Menzies here focuses on the fact that Woodward acknowledges that different agents can make different judgments about what is a cause of some event, in virtue of making different judgments about what were "serious possibilities" on that occasion. In short, we may differ in what we choose to model as a variable of the system, rather than as a fixed parameter (or not at all, to the same effect). It seems to me, however, that (as I also argued in section 5.6.2) such a disagreement isn't necessarily understood as a disagreement about what the objective causal laws governing a system are, but reflects rather a difference in the pragmatic considerations about what out to be considered as moving parts and what are fixed parts, of a system. It seems to me that the debate between Strevens, Woodward, and McCain reviewed here addresses a deeper worry about how variable relativism may undermine causal realism in Woodward's theory—a worry that it seems we can at least tentatively put to rest.

7.4. Conclusions: manipulation in Woodward's theory

Woodward's interventionist theory of causation wears the logical role of manipulations on its sleeve. The conditions for an event being an intervention articulate precisely the properties in virtue of which manipulations—when they are interventions—allow us to infer new causal information about a system. They show how an intervention establishes the direction of causation, and excludes cases of spurious correlation. This is the great advance in Woodward's theory, as compared to the earlier ones we have reviewed. In fact, we can use this theory to diagnose what is missing in the earlier manipulationist theories, as I have sometimes implicitly done, and will do more explicitly in the next chapter.

However, while the "first order" question—about what the role of manipulation is in the theory, and in virtue of what properties it can have this role—has a plain and unambiguous answer, there is a "second order" question that is more complicated. This concerns what Woodward's theory is a theory about. We might perhaps put the question more clearly: what is it that satisfies Woodward's conditions on causes, when they are satisfied? This seems to connect issues of grounds and of metaphysics—questions that Woodward has been loath

to engage with—directly to questions about the proper interpretation of the theory. I think that a comparison to the theories that interventionism has the closest relationship to may cast some further light on the issue.

7.4.1. Interventionism vs. earlier manipulationism. Manipulationism was introduced as a theory of causation when causal skepticism was a more popular view than it is today. Collingwood accepted Russell's 1912 argument that causation was not a basic feature of the physical world. Thus, it wasn't just *prima facie* unproblematic to explain or ground our notion of causation in particular features of human psychology and human choices and interests, it could seem unavoidable. The atmosphere was Humean, and as in Hume's theory, causation (specifically, its characteristic necessity or efficacy) could be seen as an artifact of particular agent beliefs and experiences. While Collingwood thought that our projection of a causal relation onto natural events, that involved no human interests or possibility of human intervention, psychologically explained a *false* belief about the existence of such a relation, von Wright's and, in particular, Huw Price's theories aim to explain rather how some *true* beliefs we have about causal relations between such natural events can depend on features humans have in virtue of being agents. This broadly Kantian view is particularly explicit in Price; the compatibilism in von Wright's account makes his position more complicated to characterize. But in all these cases, the theories are antirealist with respect to causation, at least in the sense that some part of our causal concept (such as its asymmetry or its counterfactual implications) is not grounded in a basic feature of the objective, physical world. (I don't feel confident speculating about how Gasking thought about these things, based only on the short paper reviewed in this monograph.) Again, this makes anthropocentrism the very point of their theories.

Woodward, on the other hand, gives a strong and explicit defense of causal realism in *Making Things Happen*. He says:

> Consider, for example, the hypothetical experiment in which I step in front of a speeding bus. Whether I will be injured in such an experiment does not depend, either causally or in some other way, on my beliefs or desires. (Woodward 2003, p. 119.)

Causal realism allows, and prompts, Woodward to characterize a (successful) manipulation in causal terms, rather than to take it as a theoretical primitive. But his further commitment to manipulationism leads to a theory that lacks grounds. Woodward thinks that this is not a problem: he has no interest in providing grounds for the causal relation, or in reducing causation to non-causal facts. Just as was the case with critics of earlier manipulationist accounts (Woodward was one of these critics), those who originally reviewed *Making Things Happen* tended to assume that Woodward's project was at least partially metaphysical. For example, Eric Hiddleston stated that "Woodward takes this connection with control as the metaphysical basis for an account of causation" (2005). Similar impressions were reported by Strevens (2007) and Glymour (2004). And just as was the case with Menzies's and Price's paper

(1993), I think such impressions are forgivable, given how Woodward formulates his theory there. Woodward's response has however been to disavow any metaphysical ambitions, and describe his proposal in contrast as a "functional" or "methodological" theory of causation. A comparison to the second, arguably more important, source of influence on Woodward's theory may help us understand what this could mean.

7.4.2. Interventionism vs. causal modeling. Here is one way of understanding interventions within the framework of Causal Modeling, when this is regarded as a purely formal theory. An intervention in this theory is not a possible physical event that changes a property of some individual. An intervention is just a formal operation on a theory. It transforms a model, in accordance with precise rules, into another model. The change in some variable Y that is a *theoretical* consequence of this operation is by definition the causal effect on Y, of the intervention. The primary reason for defining this operation in the theory is however not so as to provide a definition of causation. Rather, the operation precisely defines a transformation of a model such that in the new model an in-principle observable correlation between certain variables *implies* a causal relation of a certain magnitude in the original model. In this theory, the relationships between causes and interventions are stipulated in definitions. Thus, the question as to whether there really is a possible intervention on some variable in a theory or not, does not arise. There is, by definition. If the inferences endorsed by the theory are valid, and the theory is moreover sound—in the sense that whenever the theoretical premises are true of the world, the causal conclusions derived in the theory are true of the world—then the theory allows us to predict the consequences of manipulations, and explain why experiments allow us to identify causes and effects. The vast formal work in Causal Modeling theory, done by Pearl, Spirtes et al., and others over the last several decades, concerns the validity of inferences, given the theoretical definitions. As to soundness, a great benefit of the practical sciences, as compared to philosophy especially, is that there is plenty of opportunity to simply try a theory out on actual scientific problems and see how well it performs. This has also been done, in a variety of scientific disciplines. But we might be interested also in an explanation of how the formal theory relates to the world. This, I take it, would turn the formal theory into a philosophical theory of causation proper. Is this the task that Woodward takes on in *Making Things Happen*?

On the face of it, the semantic project that Woodward originally describes is a good fit for this task. Providing statements about the world as truth conditions for theoretical propositions indeed connects the theory to the world. But it's a prerequisite of this approach that we assume that we already have an understanding of the statements about the world that interpret the theory. They must thus be taken as basic in the semantic theory. As several critics have pointed out, the circularity in Woodward's definitions is thereby fatal to the semantic analysis.

Woodward has stated that he is not interested in providing something like metaphysical grounds for the causal relation, and that criticisms of his theory

have often been misplaced for this reason. But it seems to me that circularity is not just a problem for metaphysics. It is a problem for a semantic theory, and as seen in section 7.3.1, it is a problem for an epistemic theory. Circularity is, I take it, usually considered a problem for logicians and mathematicians, too, and for anyone who wants to explain something. To take an example from mathematics, we can in probability theory define conditional probability in terms of joint probability, or joint probability in terms of conditional probability. Either is a common and valid choice, but we can't do both in one and the same articulation of the theory or probability. This is not for metaphysical reasons, but just because we would then have failed to *define* either. And when this choice is made, it is often for subtle reasons. For example, when Judea Pearl introduces probability theory in *Causality*, he states that "Contrary to the traditional practice of defining conditional probabilities in terms of joint events [...], Bayesian philosophers see the conditional relationship as more basic than that of joint events..." (Pearl 2009, p. 3-4). While it makes no difference as to the theorems in the theory, then, the choice is nevertheless not wholly arbitrary. The circularity in Woodward's theory can be eliminated by taking either "direct cause" or "intervention" as basic. I.e., we could refrain from defining "direct cause" in terms of interventions, or we could remove the causal definition of interventions. The first approach seems congenial to what Woodward says in his most causally realist moments, but the latter fits better with his stated intention of "cashing out" the meaning of causal claims in terms of interventions, and with his general sympathy for manipulationism. It's thus hard to say what is the best choice of theoretical primitives, from Woodward's perspective, but it seems to me that a choice must be made, if Woodward's theory is to successfully define anything.

"Circularity" is a notion most obviously applicable to *explicit definitions*. But there are other ways of defining a concept, where circularity may not be an issue. In the next chapter I will assess Woodward's theory in light of the most well understood kinds of definition.

More recently Woodward has preferred to describe his project as "functional," and said that it is a theory about causal cognition, focused on what makes such cognition reliable and useful. What this suggests to me is that the goals of the theory may in fact be very similar to those of the formal Causal Modeling theory, as I interpret it here. According to this understanding, we could perhaps take Woodward's theoretical conditions as *rules* for causal reasoning. There is no problem of circularity under this interpretation of the conditions, as the set of cognitive rules don't metaphysically ground, or explain the nature or meaning of, anything. In fact, they look compatible with almost any type of philosophical theory of what causation is, or of what makes causal claims true. On this way of understanding Woodward's theory, the manipulationist necessary condition implies that if we think that X is a cause of Y, then we *ought to think* that there is a possible intervention on X with respect to Y, such that Y changes under this intervention—because this is a helpful way to think about causes. This would seem to explain how the necessary condition ends up being such a weak constraint in the theory. That

this functional theory of causal cognition is indeed functional is moreover not given a theoretical justification, but is justified pragmatically, by its successful application. I think this comes close to how Woodward thinks about his theory now. But this interpretation implies, I think, that we need to forget everything *Making Things Happen* explicitly claims about semantic goals or about definitions, and Woodward has not to my knowledge revised his description of the theory or its goals accordingly. This interpretation also isn't in perfect harmony with everything else Woodward says in *Making Things Happen*. For example, his original insistence that it is important to know the correct sense of "possible," in the condition that on every cause there is a possible intervention, now seems misplaced. And the claim that species membership cannot be a cause of anything, because there is no coherent notion of changing the species of an individual, seems to belong rather to a metaphysical theory of individuation of particulars, than to this functional theory of causation. After all, nothing formally prevents us from modeling species membership as a variable.

In conclusion, my impression is that whether we understand Woodward's theory as a semantic or epistemic theory, or as a functional or methodological theory of causal cognition, we must do some violence to Woodward's original formulation of it. What changes one deem necessary may indeed depend on one's metaphysical views.

The logic of Woodward's account, and some implications for its success as a theory of the meaning of causal claims, will occupy us for the brunt of the final chapter.

CHAPTER 8

The role of manipulation in theories of causation

8.1. Introduction

8.1.1. Goals and outline of arguments. The purpose of this monograph is to examine the role that manipulation has in a certain group of theories of causation, as well as to make some judgment about what role manipulation can and ought to have in our theories about causation and our relationship to it. As we have assessed the theories in the previous chapters, it has also seemed quite hard to pin down exactly what these theories are theories *about*. This chapter will scrutinize more closely whether manipulation or intervention can have the particular theoretical role of *explaining what it means to say* that something is a cause of something else. (I will take this to be closely related to giving the conditions under which something *is* a cause.) I will argue that the results of this inquiry relate directly also to the question of what a manipulationist or interventionist theory is best taken to be a theory about.

I will proceed in the following way. First, I will consider whether the older, explicitly manipulationist theories of causation can succeed in explaining what the meaning of causal claims are. I will argue that they cannot, due to an already well-known failure to provide a sufficient condition for the causal relation. I will then examine the circularities that seem to be unavoidable when the conditions in these theories are made sufficient. This takes us naturally to the interventionist treatment, the preeminent example of which is Woodward's theory.

I will argue that the interventionist theory also fails to explain the meaning of causal claims. This is due firstly to the—also familiar—way in which causation and intervention are interdefined in this theory. The circularities that result imply that Woodward's definition of direct causation doesn't succeed as an explicit definition. Secondly, I show that the conditions in Woodward's theory are also too weak to provide an *implicit* definition of direct causation. I then discuss what motivations and justifications appear available for a strengthening of the conditions that would make a definition of direct causation possible—comparing again to some views from the context of causal inference theory in science. I argue that the needed strengthening of the theory, and the particular reading of its conditions that would provide for an implicit definition, plausibly receive no support from the scientific context, and that we have very little else by way of arguments. Thus, Woodward's semantic project in *Making Things Happen* fails, at least relative to the established theories of definition. I

conclude, then, that neither manipulationist nor interventionist theories of causation can tell us what it means to say that something is a cause of something else. (And equally, they don't tell us what causes are.)

However, the fact that the interventionist theory constrains the interpretation of the causal relation in a certain way implies that, suitably modified and understood, it might support a useful *theory for causal inference under intervention*. For this reinterpreted interventionist theory to work as a theory of causal inference, we must acquire a real causal relation, either as a theoretical primitive or as a consequence of some other theory of causation. I argue, moreover, that if this is our theory of causal inference, then we need some further story about how we can come to know of some causal facts by non-inferential means, so that causal inference can get off the ground. I then sketch one such account, that is substantially informed by Woodward's definition of an intervention. Thus, I conclude that the interventionist theory is most fruitfully taken, not as a theory of causation or the meaning of causal claims, but as a theory about *interventions*. And, furthermore, that the interventionist theory, when taken in this epistemic vein, can contribute significantly to our understanding of how we come to know about some of the causal facts that hold of our world, by illuminating the unique epistemic role of manipulations.

8.1.2. Explaining causation in terms of manipulation. It has been the stated goal of the theories we have examined here—by Collingwood, Gasking, von Wright, Menzies and Price, and Woodward—to explain what it means to say that something is a cause of something else. Manipulation is at the center of these explanations (with the exception that Woodward ultimately formulates his theory in terms of interventions that don't need to be manipulations).

The idea that the meaning of causal claims can be elucidated by considerations of what happens under a manipulation shows up also in more scientifically oriented contexts, as we have seen (chapter 6). Woodward quotes Kevin Hoover's expression of a definition of causation that is "widely acknowledged" among economists: "A causes B if control of A renders B controllable. A causal relation, then, is one that is invariant to interventions in A in the sense that if someone or something can alter the value of A the change in B follows in a predictable fashion" (Hoover 1988, p. 173). The same idea can appear to be present in Judea Pearl's theory of causal inference, specifically in his operationalizations of "structural equation," "structural parameter," and of the error term in structural equations (Pearl 2009, p. 160-162). Some people who work within the *potential outcomes framework* for causal inference seemingly give manipulations a yet more prominent role in our understanding of causation. In addition to Rubin and Holland's slogan "No causation without manipulation," some insist that an effect of some factor X on an outcome Y is not *defined* in the absence of a well-defined intervention (see e.g., Hernán 2016; Broadbent et al. 2016).

In section 8.5 I will however argue that by "intervention" Pearl and Hérnan, in particular, plausibly mean by these statements something very different from what is required if the interventionist theory is to successfully *define* causation.

8.1.3. The standard theory of "meaning-explaining" definitions. The rigorous way of explaining the meaning of a term or a predicate is to *define* it. Definitions of causation are moreover abundant in the literature under consideration. When I speak of a theory of causation, below, I will normally mean a theory *qua* explanation of the meaning of causal claims, or as an explanation of what causation is, to the extent that these are closely related issues.

As to the properties of a successful explicit definition, I will rely on the standard theory as presented by Belnap (1993). This theory introduces two demands on a general "meaning-explaining definition," namely "the criterion of *eliminability* (which requires that the defined term be eliminable in favor of previously understood terms) and the criterion of *conservativeness* (which requires that the definition not only not lead to inconsistency, but not lead to anything—not involving the defined term—that was not obtainable before)" (Belnap 1993, p. 117, my emphases). These conditions are introduced to ensure that a definition gives the *whole* meaning and *only* the meaning of the defined term. Since our interest here is in examining to what extent these definitions successfully explain that which is defined, we shall focus mainly on the first part, the criterion of eliminability. Violating the second criterion, the criterion of conservativeness, means that the definition does more than provide meaning, by also implying some new facts in terms of the definiens. If this in turn means that the purported definition now can be true or false, depending on what is the case in the world, it no longer qualifies as a definition in the standard sense, but should rather be called a theory. While such a situation would strictly speaking be a problem for a proposed definition of causation, it may be less damaging with respect specifically to the goal of explaining what causation is, so I will not speak more about the conservativeness criterion here.

I will assume here that the issue of successfully explaining the meaning of a notion by providing a definition precedes any questions about realism or anti-realism with regard to that which is explained. The conditions on a successful definition are independent of whether the definiens refers to things in the objective physical world, to psychological things, to "projected" things (e.g., in Huw Price's sense, see ch. 7), or to a mixture, or something else altogether.

The main focus of the next section is theories that refer explicitly to manipulations in their definition of causation. Woodward's theory appears to rely on an intuitive idea of manipulation, especially in the early parts of *Making Things Happen*, and he does call his theory "manipulationist" (2003, "Introduction and Preview"). But his is also an interventionist theory, in that in the final formulation of the conditions on causation, it is interventions, in a technical sense, that are referred to. The interventionist attempt at a definition of causation will be the topic of the subsequent section, and the brunt of this chapter.

8.2. The manipulationist definition of causation

8.2.1. Terminology, and a first stab. In this section I will focus on those theories that give manipulations, in virtue of some properties of them, a logical role in the theory. This means von Wright's proposal, and those that come after. Collingwood's theory is, as I've mentioned, different, and I've speculated that it is partly because it arrives before Hempel's logical theory of causal explanation was published, and problems with it discussed. Thus, Collingwood's theory does not employ manipulations in the way of later proposals. I have also argued that the logic of Gasking's theory, and its use of manipulations, is radically different from the later ones. The theories I will be particularly concerned with, then, all assume that reference to manipulations in the right definition of causation can solve logical problems such as excluding cases where a correlation is due to a common causes, or establishing the direction of the causal relation. It's good to fix some terminology before we start.

By a *factor*, I mean something that might be cause or effect in a true general causal claim. It is what variables "X," "Y," "Z," ... denote in the context of causal models. These factors are determinable properties, meaning that they are properties that can take on one of several mutually exclusive values. They are moreover properties of some type of causal system, and sometimes of some *part* of the system, that can be distinguished from the system's other parts. The type of system may be, for example, a type of family, a type of person, a type of experimental setup, a type of economic system, or whatever else we have chosen to study. Examples of factors may then be, "Age of parent 1" (which takes a positive whole number as a value), "Pressure in upper cylinder" (which takes a positive real number as a value), or "Visited Crete in the past year" (which takes values 1 or 0). Since they are properties, a causal relation between factors is a relation of general causation.

An *event type* is the taking on of a value by some factor. Event types are thus denoted by "$V = v$" for some variable V and some constant v.

A *particular* (or *token*) event is an instantiation of an event type, involving some particular, and can be expressed as "$V = v$ for u," where u is an individual unit.

A *unit*, or individual, is thus an instance of a type of causal system. A row in a table of raw data typically records particular events (variable values) for one unit of observation. In any unit, every factor has one and only one value, and the observations recorded are assumed to be of instances of the same causal system (or at least sufficiently similar causal systems to allow for statistical estimation and inductive extrapolation).

A *manipulation* is here an intentional and voluntary action by an agent, involving a bodily movement that has a direct result in the external world.

A *direct result*, in turn, is the change or counterfactual difference (i.e., when the manipulation holds something fixed that would otherwise have changed) in some factor, that is due to a manipulation. For historical reasons, we do not presume that the manipulation *causes* its direct result.

An *outcome* is the value of the factor in the studied system whose causes we are interested in. In a causal model, it is usually the dependent variable of the only equation in which it occurs.

I will start by giving a "dummy version" of a manipulationist definition of causation. It does not at this first stage precisely correspond to any theory we have reviewed in the previous chapters.

MC$_1$: For every ordered pair (X, Y), X is a cause of Y *iff*

(1) X and Y are distinct,
(2) there is a practically possible manipulation M of X, that has a change in the value of X as its direct result, and
(3) Y would change under M.

By requiring that X and Y are distinct, we don't just mean that these factors are non-identical, but something stronger: they must be metaphysically, logically, and mereologically independent of each other. Otherwise, if there are any dependencies between X and Y, then they may be non-causal. We make the following assumptions. Any causal relation between X and Y is as usual relative to some relevant background conditions, that are left implicit. M is a binary variable that indicates the occurrence ($M = 1$) or non-occurrence ($M = 0$) of a manipulation, whose direct result is that X has a certain value. By a *change* in a variable, we generally mean a change in its probability distribution, with deterministic changes as a special case. For simplicity, I will speak here of values of variables, and changes in these values, and take for granted that talk of probability distributions over a variable's possible values could be substituted if needed. A change in a variable is thus defined as a pair (v_1, v_2) of distinct and temporally consecutive values of the variable. A manipulation may in general "hold fixed" the value of a variable—in which case $v_1 = v_2$ and there is no change—or it may change its value. In some definitions—for example von Wright's—it is an explicit alternative to an X-changing manipulation that M instead holds X fixed, and the corresponding condition 3 is then that Y also remains fixed. I will take this alternative as implicit here, for conciseness. The first value v_1 may moreover be the *actual* value of the variable, in a theory that focuses on actual causation, or both v_1 and v_2 may be hypothetical values, if the focus is on general causal dependencies. In either case, the causal claim in **MC$_1$** is assumed to have counterfactual implications when applicable, and to constrain what is causally possible. It thus has the modal status of a causal law. (Some manipulationist theories, such as Collingwood's and von Wright's, are explicitly concerned with explaining this modal status, by reference to some aspect of agency, but I will not engage with these issues here.)

Some of the earliest complaints about the manipulationist definition of causation concerned *circularity*. We have encountered several different types and sources of circularity in these treatments, all of which have appeared avoidable in one way or another. A major focus in this chapter is therefore to disentangle the issue of circularity in manipulationist treatments of causation. I

will connect this concern to the goal of explanation, by relating it directly to the standard theory of definition. The important thing for us, then, is that *circularity defeats eliminability*. Assume that in addition to **MC$_1$**, we have a definition of "manipulation of _" that gives as a necessary condition that the manipulation is a cause of its direct result. Substituting this definition in the definiens of **MC$_1$** leads to an explicit circularity. This has the consequence that the definiendum cannot be eliminated in contexts in which it occurs, since the substituted expression also contains the definiendum. This particular situation was the target of one of the earliest criticisms of manipulationist definitions of causation, and I will dub it "circularity$_1$." In addition to this circularity, early critics saw another one, that we get if we in **MC$_1$** replace "Y would change under M" with "M would change Y." That is to say, we substitute a causal, transitive "change" in condition 3 for the current non-causal intransitive use of the word. When the condition that the changer is a cause of the change is imposed, then again eliminability is violated. I will call this "circularity$_2$."

I have argued in the previous chapters that the definitions by Collingwood, Gasking, and von Wright in fact do not suffer from circularity$_1$, because their theories imply that the manipulation does not cause its direct result. Menzies and Price aim for something similar by suggesting that our personal "bringing about" of things is prior to "causing" in our acquisition of those concepts. Moreover, it has seemed to be a simple matter to avoid circularity$_2$ in a definition of causation, by formulating condition 3 in the non-causal way, as shown in **MC$_1$**. However, I will say more about some consequences of doing this, below.

Leaving the issue of circularity for a moment, all advocates of manipulationism that we have encountered would agree that **MC$_1$** is not an adequate definition of causation. Collingwood would consider the condition on causation that is stated in **MC$_1$** to be insufficient, because the agent for which X is a cause of Y must, according to him, also have a vested interest in controlling Y by way of X. But the more mainstream objection would be that the condition is not *necessary*. This is because the partial condition 2 is not necessary. That is, the cause X is not itself necessarily practically manipulable, it must merely be *related* in some specific way to something practically manipulable. This, then, concerns the manipulationist necessary condition, that I took to characterize manipulationist theories, in chapter 1. The necessary condition implied by **MC$_1$** states that, if X is a cause, then X is manipulable. The modified necessary condition now states instead that, if X is a cause, then X is manipulable or X is related in a certain way to something manipulable.

How X must be related to something manipulable differs between different formulations of the theory. Gasking says that X must be of a "sort" that can be manipulated. Von Wright claims in addition that X, when it is a cause, is composed of parts that are of some manipulable sort. Menzies and Price specify the relation further by stating that X shares some non-causal properties with manipulable things. I will generalize this to a relation R that X must have to something practically manipulable, for it to be a cause, and leave the definition of R to the respective theories. It is clear that Y, too, must be related in some corresponding way to the effect of the manipulable thing that X is related to.

E.g., the natural cooling of the atmosphere (the unmanipulable cause) is to rain (the effect) as the artificial cooling of a glass jar containing water vapor (the related manipulable cause) is to the appearance of water droplets in the jar (the related effect). We have, then, an amended version of definition.

MC$_2$: For every ordered pair (X, Y), X is a cause of Y *iff*

(a_i) X and Y are distinct,
(a_{ii}) there is a practically possible manipulation M of X,
that has a change in the value of X as its direct result, and
(a_{iii}) Y would change under M,

or

(b_i) there is an ordered pair (X', Y') such that
(b_{ii}) it satisfies conditions (a), and
(b_{iii}) $R(X, X')$ and $R(Y, Y')$.

R here stands for "has a part of the same sort as," "has a part that shares certain non-causal properties with," or something along these lines. (It may not be a proper part.) The adequacy of **MC$_2$** cannot be assessed unless we know more about R. For example, if R is understood in Menzies and Price's sense (referring to shared non-causal properties), have we any reason to think that it is not in virtue of having these properties that X and Y are causally related, in which case the account risks collapsing into a governing-law theory of causation? That is to say, there is a strong suggestion that some causal law is what connects these properties as cause and effect, rather than truths about what would happen under a manipulation. Moreover, a precise formulation of this definition invites new questions. For example, must *every* part of X be of the same sort as *something* that is a practically manipulable cause of something that is of the same sort as some part of Y? If so, then **MC$_2$** requires additional conditions. (Because, while these unmanipulated parts come together in X, there may be no X' in which all the related manipulable parts come together. Likewise for Y.) Another potential issue concerns the *interaction* of the parts of X, that may lead to different effects compared to the parts when they occur separately. That we need to know more about R to evaluate this part of the theory is the extent of what I will say about the necessary condition in relation to these explicitly manipulationist theories. I think the more interesting issue lies in the condition they take to be *sufficient* for causation. (I will however return to the necessary condition as it occurs in interventionist theories.)

8.2.2. The insufficiency of the manipulationist theory of causation. Here I argue that the condition that manipulationism gives for the presence of a causal relation cannot be *sufficient*. This means that the problem now is not that some causes are not practically manipulable, or not related in the right way to such manipulable things, but that some manipulations of X are such that even if Y changes under this manipulation, X may not be a cause of

Y. The conclusions I present here are not new, I aim to summarize and clarify insights that have at the very least been implicit, and sometimes explicit, in the existing literature on manipulationism. In particular, Woodward makes a point that is closely related to the one I make here, when discussing the prospects of a reductive manipulationist theory of causation (2003, p. 28). However, I emphasize this situation to greater extent, because I will subsequently argue that, as a consequence of it, Woodward's semantic goals are also defeated.

The reason why $\mathbf{MC_2}$ cannot be giving a sufficient condition for the presence of a causal relation between X and Y should thus be familiar at this point. If the manipulation of X is understood as an intentional action involving a bodily movement, the direct result of which is a change in X's value, then this event may take place, and the value of Y change, even though X is not a cause of Y. This may happen under one of the following two circumstances.

(1) The manipulation event $M = 1$ has a common cause with the change in Y.
(2) The manipulation event $M = 1$ is a common cause of the change in X and the change in Y.

As $\mathbf{MC_2}$ does not exclude these types of situations, the theory does not state a sufficient condition for X to be a cause of Y.

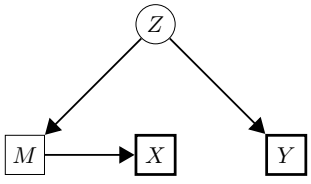

FIGURE 8.2.1. Z is a common cause of the manipulation M and the outcome Y.

The following imagined situations may illustrate these possibilities. If I decide to press a button ($M = 1$ sets $X = 1$) just when a light on the wall comes on ($Z = 1$), then the simultaneous toot of a horn ($Y = 1$) may be explained by the horn being connected to my button by one circuit, or by the horn being connected to the light by a different circuit. In the latter case, my action is caused by the activation of the circuit that turns on the light, which also turns on the horn. That is to say, by intentionally cueing my manipulation to an external event, the possibility is introduced that the manipulation and the observed outcome have a common cause, as seen in figure 8.2.1. This is a failure of the causal exogeneity of the manipulation to the manipulated system. Our decisions can be similarly affected by external events in all sorts of ways that are unknown to us at the time.

In a less direct, but more realistic, example, I may be selecting my sample from an already biased set, and compare the outcome to the population at large. Thus, if a treatment was given exclusively to volunteers, who volunteered because they know they have a sturdy physical constitution, then a significant positive outcome, compared to untreated persons in the whole population, may be explained by the sturdy physical constitution of those treated, which causally affected both the probability of being selected for treatment, and the probability

of a positive outcome. (And this is why we need properly randomized and controlled trials.)

And the sufficiency of $\mathbf{MC_2}$ can be violated for reasons other than failure of exogeneity. For example, I may test a hangover treatment by giving a group of people who are hung over a newly developed hangover pill ($M = 1$ sets $X = 1$) and a glass of water, and compare rate of recovery ($Y = 1$) to a hungover group that receives nothing. While there may be a significant effect on the treatment group, this could be caused by drinking a glass of water, rather than ingesting the pill. (I.e., $M = 1$ also causes $Y = 1$ independently of X.) In this sort of case, the problem is that the manipulation causes the effect independently of the putative cause. The manipulation is thus itself a confounding common cause of the treatment and the outcome, as seen in figure 8.2.2. (And this is one reason why we need placebos.) Henceforth, I will call manipulations that introduce problems of any of these two types "*confounding manipulations*."

It may be worth noting before we move on that we are using the expression "common cause" as a term of art in these contexts. In figure 8.2.1, Z is a common cause of M and Y in the way we mean when we say "common cause." But, strictly speaking, Z is also a cause that M and X have in common. When we say below that X and Y may not have a common cause, and aim by that to exclude situations of type 2, we don't intend to exclude the possibility that M is a common cause of X and Y in virtue just of being a cause of X, and *thereby* a more remote cause also of Y. "Common cause," then, refers to this particular fork structure, which, we may also note in passing, means that this concept requires a notion of *causal paths*.

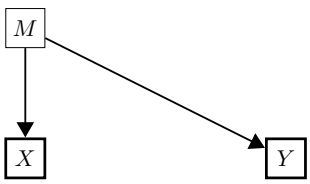

FIGURE 8.2.2. The manipulation M is a common cause of X and the outcome Y.

More has been assumed about manipulations in these theories, than that they are intentional actions involving a bodily movement, that result in some change in the external world. Von Wright and Huw Price have been the most articulate on this point, von Wright claiming that it is a category mistake to think that *actions have causes*, and Price adopting the assumption, that he attributes to Ramsey, which says that actions have no causal history beyond the agent herself or her intentions. Under the assumption that actions have no causes external to the agent, situations of type 1 above are excluded. But since such situations *are* possible, the assumption cannot hold for manipulations in general. The reference to manipulation in $\mathbf{MC_2}$ must therefore be qualified, and changed to "free manipulation" or some such, where this is taken to denote a class of manipulations such that they have no external causes—or they at least do not share a cause with the accompanying change in Y. (I.e., the manipulation is causally exogenous at least relative to the outcome.) Note that "free" can then not mean–as it has often seemed to do—merely "voluntary." For example, in the button-horn scenario described above, the button press is an example of a voluntary action, that is not free in the special causal sense.

Constraining the condition in **MC₂** to (causally) "free" manipulations does not address possibilities of type 2, above. A common informal characterization of a condition that does exclude this possibility is that the manipulation that results in X taking on a certain value must be "surgical" (e.g.: Pearl 2009, p. 224; Woodward 2003, p. 130). What this means is that the manipulation must not have any *other* causal consequences than X taking on this value, or at least it cannot have any consequences that cause the change in Y independently of the change in X. Surgicalness thereby excludes situations of type 2, where the manipulation is a common cause of X and Y.

We can see, then, that a manipulationist definition that provides a sufficient condition for X to be a cause of Y, must require that the manipulation is "free" in the sense of causally exogenous at least relative to the outcome, and that it is "surgical" in the sense of not having some effect that affects the outcome Y independently of X. Adding these conditions on the manipulation M of X, we get the manipulationist definition **MC₃**. **MC₃** goes beyond what has been suggested in purely manipulationist theories, and approaches the *interventionist* definition of causation.

MC$_3$: For every ordered pair (X, Y), X is a cause of Y *iff*

(a_i) X and Y are distinct,
(a_{ii}) there is a practically possible manipulation M of X, that has a change in the value of X as its direct result,
(a_{iii}) M is not a cause of Y unless by way of X,
(a_{iv}) nothing is a cause of M and also of Y independently of M, and
(a_v) Y would change under M,

or

(b_i) there is an ordered pair (X', Y') such that
(b_{ii}) it satisfies conditions (a), and
(b_{iii}) $R(X, X')$ and $R(Y, Y')$.

The new conditions a_{iii} and a_{iv} are the surgicalness and exogeneity conditions, and they employ the "_ is a cause of _" relation. This new circularity, that is introduced in order to exclude confounding manipulations, I will call "circularity$_3$." (It is new in our discussion here—as I noted above, Woodward has talked about it.)

Note that there is a way of stating the manipulationist theory so that the surgicalness and exogeneity conditions are *not* required. Consider Collingwood's formulation, below.

> A cause is an event or state of things which it is in our power to produce or prevent, and by producing or preventing which we can produce or prevent that whose cause it is said to be. (Collingwood 1940, p. 296-297.)

The conditions explicitly state that the manipulation *produces* or *prevents* the outcome, by way of its direct result. Thus, the possibility of a confounding manipulation, in the sense given above, is excluded. However, assuming (as is usual) that "produce" and "prevent" are causal terms, Collingwood's definition clearly suffers from circularity$_2$ and is naturally perceived as being blatantly question-begging. On the other hand, eliminating this circularity in favor of a non-causal "the outcome would change if the manipulation occurred" forces us to add instead the unavoidably causal exogeneity and surgicalness conditions. It appears then, that sufficiency of the theory's conditions implies either circularity$_2$ or circularity$_3$. Moreover, since either insufficiency or circularity is implied, there is no definition of causation in terms of what happens under a manipulation that satisfies the conditions imposed by the standard theory of definitions.

We can thus state our conclusion, as regards a theory that aims to provide the meaning of causal claims in terms of what happens under a manipulation by a human or other agent, in the following way. Assuming that the standard theory of definition is the right account of what it means to provide the meaning of a concept or a type of claim, a theory succeeds in doing this only when it is such that the defined term can be replaced everywhere by a statement expressed in different terms, that we already understand. This is the eliminability criterion on definitions. In a manipulationist definition that contains conditions a_{iii} and a_{iv} in $\mathbf{MC_3}$, this isn't possible, since every substitution of the definiens for "_ is a cause of _" introduces new instances of "_ is a cause of _" from these conditions. I.e., the definition is circular and eliminability is violated. If, on the other hand, the theory does *not* contain conditions a_{iii} and a_{iv}, then the theory will identify some non-causes as causes, namely X in those cases where Y would change under the manipulation M on X, but where M is a confounding manipulation, suffering from at least one of the problems shown in figures 8.2.1 and 8.2.2. Thus, the theory's conditions have in this case been shown to be insufficient.

Here we have only considered to what extent a traditional manipulationist theory can satisfy the immediate requirements of the standard theory of explicit definitions. This doesn't exhaust the question if and how a theory of this sort can succeed in defining a causal relation, for example *implicitly*. That question will be treated in some detail in section 8.4.4, in relation to Woodward's interventionist account.

We can make the further observation at this point that, if we make explicit the commonplace assumption that a manipulation is an event that causes its direct result, then specific reference to manipulation in $\mathbf{MC_3}$'s condition a_{ii} becomes *redundant*. The sufficient condition remains sufficient if "possible event `E` that is a cause of X" is substituted for "practically possible manipulation M of X." It seems to make no sense to talk of "practically possible" events in general, so this qualification must also go. This in turn leaves it open what sort of possibility `E` must have according to the new formulation. We are thus approaching a discussion of the interventionist definition of causation, which

differs from **MC$_3$** mainly by not referring explicitly to manipulation, and by its different take on the manipulationist necessary condition.

If manipulations are taken to not cause their direct result, then explicit reference to manipulation can do some substantial work in the definition—but this appears to be a highly unnatural and highly unusual thing to believe. (It is not clear to me to what extent this implication was really intended especially in Gasking's proposal.) Assuming instead, then, that the direct results of manipulations *are* caused by the manipulation event, then since the minimal sufficient conditions for the presence of a causal relation do not, under this assumption, include reference to manipulation, reference to manipulation in such theories, if it accomplishes anything, must accomplish *something else*. At the broadest level, reference to manipulation connects causation to something with which we are all already familiar: our agency. In the scientific context, it connects causation specifically to experimental practices. This connection can therefore appear illuminating or informative in some way. But our argument here shows, I think, that by making this connection in the *definition* of causation, or in a theory aimed at explaining what causation *is* in general, manipulationist theories misplaces it. I will extend this conclusion to the interventionist treatment of causation in the next section. In a later section I will suggest a better place for this connection between causation and manipulation, that I think affirms the intuitions driving the manipulationist approach.

8.3. The interventionist definition of causation

8.3.1. Repeating the conditions. James Woodward aimed, in *Making Things Happen*, to provide a theory of the meaning of those causal claims that he thought Judea Pearl had taken as theoretically basic in *Causality*, which were claims about "direct" causation. From that definition, he would then acquire a definition of causation in general (what he calls "contributing causation"). While this was the clearly stated goal, there was also a focus on an analysis that makes this concept *useful* in causal cognition, an aspect that he has further emphasized in later writings. (See the previous chapter.) The project was thus reminiscent of what Carnap called "explication," and Belnap calls "analysis," but perhaps with a more directly practical focus. Woodward proceeded in *Making Things Happen* to give a definition of direct and contributing causation in terms of what happens under an intervention, and then a causal definition of interventions, in turn. In this section I will repeat Woodward's theory, and add some clarificatory remarks, according to how I understand it. I will go on to argue that the interventionist theory cannot be said to provide a definition of causation, according to the common views of how this is supposed to work, and that it therefore cannot supply the meaning of causal claims in a well-understood way.

I will focus on assessing Woodward's definition of causation in light of traditional theories of definition. But, judging from some things Woodward has said more recently, it may be that he now does not mean for his theory to provide the meaning of causal claims in the way that he indicates in *Making Things Happen*. I will say a little about this toward the end.

I will begin by repeating Woodward's definitions of causation and intervention here, for convenience. (Formatting and emphases are mine.)

> **$\mathbf{M_1}$ (direct causation):** A necessary and sufficient condition for X to be a (type-level) *direct cause* of Y with respect to a variable set V is that there be a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all other variables Z_i in V.
>
> **$\mathbf{M_2}$ (contributing causation):** A necessary and sufficient condition for X to be a (type-level) *contributing cause* of Y with respect to variable set V is that
>
> *(i)* there be a directed path from X to Y such that each link in this path is a direct causal relationship; that is, a set of variables $Z_1 \ldots Z_n$ 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 $\ldots Z_n$, which is a direct cause of Y, and that
>
> *(ii)* there be some intervention on X that will change Y when all other variables in V that are not on this path are fixed at some value. If there is only one path P from X to Y or if the only alternative path from X to Y besides P contains no intermediate variables (i.e., is direct), then X is a contributing cause of Y as long as there is some intervention on X that will change the value of Y, for some values of the other variables in V. (Woodward 2003, p. 59.)

As discussed in the previous chapter, the "directness" of the causal relation in $\mathbf{M_1}$ is relative to a certain set of variables (V): a causal relation from X to Y is direct only if V does not contain any variables denoting causes that are intermediate between X and Y. $\mathbf{M_2}$ makes contributing causation—which includes both direct and indirect causation—relative to a set of variables as well. But we have also noted that Woodward later gives us reason to add the following definition of *contributing causation simpliciter* to his theory (see, again: Strevens 2007; Woodward 2008).

> **CCS:** X is a *contributing cause simpliciter* of Y *if and only if* there exists a variable set V such that X is a contributing cause of Y with respect to V.

Woodward adds the following definition of an *intervention variable*.

> **IV:** I is an intervention variable for X with respect to Y *if and only if*
>
> *I1.* I causes X.

> *I2.* 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.
>
> *I3.* Any directed path from I to Y goes through X. That is, I does not directly cause Y and is not a cause of any causes of Y that are distinct from X except, of course, for those causes of Y, if any, that are built into the I-X-Y connection itself; that is, except for (a) any causes of Y that are effects of X (i.e., variables that are causally between X and Y) and (b) any causes of Y that are between I and X and have no effect on Y independently of X.
>
> *I4.* I is (statistically) independent of any variable Z that causes Y and that is on a directed path that does not go through X. (Woodward 2003, p. 98.)

Woodward then defines an intervention (event) in the following way (2003, p. 98).

> **IN:** 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.

8.3.2. Explanations and disambiguations. Due to the complexity of Woodward's theory, it's not feasible to give it a summary formulation, such as $\mathbf{MC_3}$ in the previous section, but several things can be explained and clarified. I should make clear that what I say below represents my understanding of some of the details of Woodward's theory, and that other ways to disambiguate certain things may well be possible. But choices must be made, and these are my best suggestions.

To begin, the fact that some of Woodward's definitions are relative to a set of variables is a major difference to the earlier proposals, and may be a source of some confusion. Woodward cleared up some of that confusion in his interchange with Strevens, that I reviewed in the previous chapter. Firstly, the definitions of "direct cause" and "contributing cause" are clearly relative to a set of variables. These, I take it, are the variables included in a causal model, of the type used in causal modeling theory, by Pearl and others. Woodward explained that the definition of an intervention (variable and event) is *not* relative to a set of variables. One way of explicating this would be to take the use of "cause" in **IV** in particular to refer to a contributing cause simpliciter, as defined in **CCS**. As I understand things, the conditions for "direct cause" and for "contributing cause" are equivalent under existential quantification over possible variable sets. A quick proof goes as follows. If there exists a set of variables relative to which X is a direct cause of Y, then it follows immediately that there exists a set of variables relative to which X is a contributing cause of Y, since direct causes are contributing causes. In the other direction, if there exists a set V of variables relative to which X is a contributing cause of Y,

then there exists a set of variables relative to which X is a direct cause of Y, namely V minus any variables that are intermediate between X and Y. This would then seem to interpret Woodward's use of "cause" in **IV**.

This also means that there is an ambiguity where Woodward quantifies over variables, in for example **M** on the one hand and in *I2* and *I4* on the other. In **M**, this is a quantification over the variables in V, while in the latter cases—specifically, "All the other variables" in *I2*, "Any direct path" in *I3*, and "any variable Z" in *I4*—it is a quantification over all of the objectively existing factors in the system, whether they are modeled by variables in V or not.

Woodward's definition of an intervention event can seem confusing in one respect. That definition states that the intervention $I = k$ on X must be an *actual* cause of X, in accordance with Woodward's definition **AC** of an actual cause (Woodward 2003, p. 76). Taken literally, every possible intervention is then an *actual* event, by definition. This is clearly not Woodward's intention. Woodward suggests a non-literal understanding of this condition in *Making Things Happen*. He proposes that, when the conditions in **IN** are not satisfied by any actual event, we should treat the definition as a "regulative ideal." That is to say, when considering the effect of X on Y, in situations where there is no actual intervention on X, "we should think of ourselves as trying to determine what would happen in an ideal hypothetical experiment in which X is manipulated in such a way that the conditions in **IN** *are* satisfied" (2003, p. 114). Thus, it seems as though we ought to read Woodward's theory as requiring that the intervention on X is a type-level event cause of X, the occurrence of which may be a mere possibility. This possibility is represented by a possible value k of an intervention variable I, in that for $I = k$, condition *I2* is satisfied in the system (i.e., the variable intervened on is disconnected from all of its other causes). Just as we might ask about the modal strength of an explicit possibility operator, we can ask in what space of possibilities I may have this value. Thus, questions about the sense and strength of "possible intervention" in Woodward's theory relate directly to I's domain of possible values.

Two things need to be said about the condition in $\mathbf{M_1}$ and $\mathbf{M_2}$ stating that there is a change in the outcome variable Y, under the intervention on X, when some other things are held fixed. First, I take "change in Y" to mean a change in the *probability distribution* of Y. This makes sense of the "fixing interventions," used in **M**: if causation is defined in terms of a literal change in the value of a variable, then it's hard to understand how an intervention that fixes that value is to be understood, but the change in the probability distribution of Y may fix Y's value to k by making $P(Y = k) = 1$. That the change is in Y's probability distribution suggests as well that what is held fixed in **M** is not exogenous causal influences (i.e., the noise variables U): the precise value of Y can vary between occasions when exactly the same set of (type level) intervention events occur. What fixing does in **M**, then, is just making sure that if there are multiple directed paths between X and Y, whose causal influences on Y ultimately cancel out, then alternative paths are blocked by holding some intermediate cause located on them fixed. The second thing

that must be said of "change" in the theory is that it's an intrinsically causal notion. Why, and what this means exactly, is explained in section 8.4.4 on implicit definition, below.

M₁ defines the basic concept of "direct cause" in Woodward's theory of causation. As I understand it, this is intended as an interpretation of, among other things, the arrows in Pearl's causal diagrams. "Direct cause" occurs in definition **M₂** of "contributing cause." Specifically, "direct cause" occurs in the definition of "directed path" that is implied in **M₂**. "Direct cause" does not occur in definition **IV** of an intervention variable, but it's implied in each use of "cause" or "directed path" in those conditions (by way of **CCS**). Consequently, while "direct cause" does not occur in Woodward's definition of "direct cause" itself, it is implied by the occurrences of "intervention" and "holds fixed."

We may note in passing that Woodward states the theory in *Making Things Happen* in such a way that it exhibits circularity$_2$, by saying that the intervention *changes* the outcome (or its probability distribution). This is a transitive, causal use of "change." Woodward amends this in later formulations, where he says rather that the outcome would change under the intervention (e.g., Woodward 2014b, p. 697). I take this to be synonymous with that the change in the outcome coincides with the intervention. Thus circularity$_2$ is avoided on the grammatical level. However, as I mentioned above, and will discuss further below, Woodward's notion of change here is still causal.

8.4. Does Woodward's theory define causation?

8.4.1. Circularity and explicit definition. We noted above that a proposed explicit definition which employs the definiendum in its definiens fails to satisfy the eliminability criterion on definitions. Whether Woodward's definition of causation is circular in this way or not depends on whether we regard it as an explicit definition and, if so, what the definiendum is taken to be. I have taken for granted that Woodward's definition **M₁** contains an implicit quantifier over the variables X, Y, and V, and that what is being defined is the three-place predicate "_ is a direct cause of _ relative to _." This is required for the causal relation to acquire an extension over the whole domain. Occurrences of this predicate in the definiens of **M₁** now constitutes a circularity, in virtue of which the eliminability criterion is violated. There are no such explicit occurrences, but they are implied by the occurrences of "intervention" and "holds fixed," as pointed out above. A difference in presentation between Woodward's theory and **MC₃** that is related to the difference in complexity, is that in **MC₃** the conditions on an intervention (a_{ii}–a_{iv}) are an explicit part of the definition of causation. We could, in principle, imagine a formalization of Woodward's theory (although this is quite complicated to produce and might involve a fair amount of new theoretical development work), such that each occurrence of "intervention" and "hold fixed" (and the implied "directed path") in **M₁** can be recursively expanded into their definiens, as these are given in their respective definitions. In this "reduced" formulation of **M₁**, the circularities would be explicit, just as in **MC₃**. Below, I will call this reduced formulation

of the theory, where interdefinitions of terms are replaced by explicit circularities, "T^W." This treatment would also make the methodological and epistemic regresses described by Baumgartner easier to spot (Baumgartner 2009).

It might be useful to consider an alternative to this reduction procedure. We could, I assume, give the definition of an *intervention* the same treatment. That is to say, we would eliminate every occurrence of "cause" and "direct path" in **IV** in favor of their definiens. This would result in a definition of "intervention" explicitly in terms of what happens under interventions. Maybe this would be more in line with some sense of "cashing out the meaning of causal claims in terms of what happens under interventions." But I take the first alternative to better capture the idea of a theory of causation. Either way, it seems to me that we are allowed, on purely formal grounds, to reduce the theory in one of these ways. To insist otherwise—to claim, for example, that we ought not to eliminate mention of "intervention" in the definition of "direct cause"—seems to imply that the term is doing more work there than what is made explicit in its definition, and we would want to know what that is. This will matter somewhat in the section on "connective analysis," below.

Thus, if we insist on interpreting Woodward's definition of causation as an attempt at providing an explicit definition subject to the standard conditions, then there seems to be no room for doubt as to its circularity. That the circularity is "vicious," from the perspective of explaining the meaning of causal claims, I then take to follow from the fact that the eliminability criterion is violated, as this entails that explicit definition fails.

This discussion has focused on syntax, i.e., the way the definitions and their conditions have been stated in the theory, and the implied circularities. That the theory as stated doesn't constitute a successful explicit definition of direct causation doesn't however settle the question as to whether the theory ultimately succeeds in defining this relation. That general question is answered in section 8.4.4 on implicit definition, below. But before we get to implicit definition, it's useful to discuss the theory in relation to two other notions of definition: P. F. Strawson's "connective analysis" and inductive definition.

8.4.2. Connective analysis. Woodward has to my knowledge not explicitly mentioned P. F. Strawson's notion of *connective analysis* in relation to his own theory. But he does at times characterize his project in *Making Things Happen* in ways that point our thoughts in that direction. This happens in particular when Woodward speaks of causation as a concept to be located in a "circle of concepts" that also includes such ones as "law," "explanation," and "intervention" (2003, ch. 1; 2011, p. 28). Woodward argues that establishing interconnections between these concepts can be illuminating, even if the account is not reductive. This is clearly true, in the particular sense that it can provide new information about the interrelationships between the concepts, at least if we *already have* some understanding of what at least some of these are concepts of. To wit, explaining that it takes at least four bobels and an anti-bobel to make up a strang, and that something is a bobel if and only if it's at most one fifth of a strang, can provide a new understanding of the structural relationships between bobels and strangs. But if we have no idea what a bobel,

or a strang, is, then the explanation fails to explain anything, and adding more intricacies to the structure will not help.

Strawson's connective analyses are not essentially circular, although he thinks that circularity is not necessarily a problem for them, in the way they are for reductive analyses. He asks us to imagine a

> model of an elaborate network, a system, of connected items, concepts, such that the function of each item, each concept, could, from the philosophical point of view, be properly understood only by grasping its connections with the others, its place in the system—perhaps better still, the picture of a set of interlocking systems of such a kind. If this becomes our model, then there will be no reason to be worried if, in the process of tracing connections from one point to another of the network, we find ourselves returning to, or passing through, our starting-point. (Strawson 1992, p. 19.)

This seems true only if we have some prior understanding of a sufficient number of the concepts involved, and only when these are situated in the "right" places in the network. Strawson recognizes that circularity can be a problem also on the connectivist view of conceptual analysis, if the "circles are too small and we move in them unawares, thinking we have established a revealing connection when we have not" (1992, p. 20).

In the case of Woodward's theory, "intervention" has a stipulated, technical definition that is supplied by the theory itself. It doesn't, and isn't meant to, correspond to our pre-theoretical notion of manipulation, except in certain special cases. It therefore seems perfectly unobjectionable to eliminate it in $\mathbf{M_1}$, in favor of its definition, as argued in the previous section. This, I think, reveals the circularity to be not just tight, but immediate, and the definition therefore a poor fit for such a connective analysis.

8.4.3. Inductive definition. Interestingly, while Woodward has pointed out that there can be no reductive manipulationist theory of causation, precisely because of the unavoidable circularities I have highlighted in this chapter, he has also argued that his definition of causation is *not* viciously circular. On the first point, he says:

> [A]ttempts to analyze causation in terms of manipulation turn out to be "circular" not just in the obvious sense that for an action or event I to constitute a manipulation of a variable X, there must be a causal relationship between I and X, but in other, more subtle and interesting ways as well: for I to qualify as a manipulation in the sense relevant to understanding causation, I must not just be causally related to X but must be an event or process with a very special kind of causal structure, and to characterize this structure we must make extensive use of causal notions. (Woodward 2003, p. 28.)

But, anticipating a circularity objection regarding the definition of causation he proposes, he says in *Making Things Happen* (with respect to the definition of intervention, as part of the definition of direct causation):

> [I]t is [...] crucially important to understand that [the definition of intervention is] not viciously circular in the sense that the characterization of an intervention on X with respect to Y itself makes reference to the presence or absence of a causal relationship between X and Y. The causal information required to characterize the notion of intervention on X with respect to Y is information about the causal relationship between the intervention variable I and X, information about whether there are other causes of Y that are correlated with I, information about whether there is a causal route from I to Y that does not go through X and so on, but not information about the presence or absence of a causal relationship between X and Y. (Woodward 2003, p. 104-105.)

Woodward's defense here might suggest that what he is defining is "X is a direct cause of Y relative to V," rather than the predicate "_ is a direct cause of _ relative to _." The former expression indeed does not occur (or is implied) in the conditions for "direct cause," so in that case there is no circularity. As we noted above, this cannot reasonably be what is being defined in the theory. "X," "Y," and "V" are not interpreted in that expression—they are certainly not individual terms with definite referents. But it's possible to understand Woodward also in a different way, as suggesting that the theory be understood as providing an *inductive* (or recursive) definition.[1] Relatedly, in section 7.3.1 I quoted Clark Glymour calling Woodward's definition of causation "ill-founded." We might take this as a claim about the theory viewed as an inductive definition of causation.

An inductive definition defines membership in a set (i.e., the extension or intension of a predicate) in terms of other members of that same set. As such, inductive definitions can be said to be "circular" in a special sense, by design. But membership for some element is not defined in terms of that *same* element's membership in the set, which corresponds to Woodward's quoted explanation of how his theory is not circular in that particular sense.

We can illustrate inductive definition by the definition of a *term* in a first order language. Let T be the set of terms in the language. S can be the set of all individual constants and variables, and is assumed defined, and f is any n-place function symbol in the language.

(1) If $s \in S$ then $s \in T$.
(2) If $t_1, t_2, \ldots, t_n \in T$ then $f(t_1, t_2, \ldots, t_n) \in T$.

(2) introduces a circularity, in the sense that membership in T occurs as a condition in the definiens. Without the independently sufficient condition (1),

[1] I'm indebted to Sebastian Lutz for suggesting this interpretation of Woodward's explanation.

therefore, membership in T couldn't be evaluated for any individual. The definition would thus be ill-founded in the absence of the *base case* (1).

We can examine a simplified expression of Woodward's theory in a corresponding way. Let C be the relation to be defined, that is to say the causal relation.

$$
\begin{array}{ll}
\mathbf{W^{IND}}\mathbf{:} & \forall xy(C(x,y) \leftrightarrow \exists z(\\
\textit{(i)} & z \neq x \wedge C(z,x) \\
 & \wedge \\
\textit{(ii)} & P(x,y,z) \\
 &))
\end{array}
$$

Under the intended interpretation, condition *(i)* states that for x to be a cause of y, something z (distinct from x) must be a cause of x. This something is the possible intervention on x, required by Woodward's theory. P stands in the place of the rest of the conditions on interventions and on the causal relation in the theory. (P contains further causal conditions, but we don't need to consider those here.) To evaluate these conditions for some factors $x = R$, $y = S$, $z = T$, we then need to know if there exists a factor z' such that $C(z', T)$, and so on, *ad infinitum*. The definition is thus ill-formed, in that the conditions aren't evaluable for any pair of factors. The ill-formedness of the theory viewed as an inductive definition is thus the precise analogue of the infinite regresses identified by Baumgartner, fatal to the theory under the expectation that it provides a method sufficient for identifying causes. What the theory lacks, viewed as an inductive definition, is a base case. E.g.:

$$\mathbf{BASE:}\quad C(T,R)$$

For the same values for x, y, and z as above, and assuming that $P(R,S,T)$ holds, it now follows that $C(R,S)$. Thus, just as the adequacy of Woodward's theory viewed as a *method* for causal inference depends on the availability of some causal information that is *not* acquired by this method, the adequacy of the theory viewed as an inductive definition of causation depends on something being a cause of something else, and *not* in virtue of what happens under an intervention (since that is the inductive step of the definition).

Perhaps the basic cases of causation are provided by our *best scientific theories*? We have reason to think that this proposal isn't compatible with Woodward's goals. If the basic cases of causation are not cases of causation in virtue of what happens under an intervention, then it's not at all clear how the theory fulfills the goal of cashing out the meaning of causal claims in terms of what happens under an intervention. Woodward seems to reject the existence of additional independently sufficient conditions for causation for just this reason, when he justifies the necessary condition in his theory:

> If there are facts about what would happen to Y under hypothetical interventions on X that are sufficient for X to cause Y but there are no such facts that are necessary for X to cause Y, we would then face the possibility that there is some other

> set of conditions, having nothing to do with facts about what would happen under manipulation of X, that are also sufficient for X to cause Y and puzzling questions about the relationship between these two sets of conditions and why they are both relevant to causation. (Woodward 2003, p. 60-61.)

Indeed, if T is a cause of R, as in the base case above, and *not* in virtue of what would happen to R under an intervention on T, then we might suspect that whatever the right explanation is for *this* instance, all the other instances of causation have the same explanation—perhaps in terms rather of, say, physical interactions (or perhaps the causal relation is taken as a primitive)—even if certain facts about what would happen under an intervention *also* hold true. The idea that causal claims have been interpreted in terms of correlation under intervention has then been effectively undermined. Woodward goes on to say:

> By providing both necessary and sufficient conditions for causation, M and TC give us a way of fully capturing or cashing out the content of causal claims in terms of facts about what would happen under interventions. (Ibid.)

But the present section argues that this isn't true, at least under the ordinary understanding of what it means to interpret a type of claim.

8.4.4. Implicit definition.

8.4.4.1. *General condition for implicit definition.* Does Woodward's theory *implicitly* define causation? This is perhaps connected to the "Euclidean" or axiomatic reading of Woodward's theory, that has been advocated by Clark Glymour (without any overt claims to the effect that an implicit definition as such would then be successful) and also discussed more critically by Michael Strevens (Glymour 2004; Strevens 2008). On this view, Woodward's account of direct causation would be understood as a theory that states axiomatically certain relations between causation, intervention, correlation, and perhaps other things, and in such a way that a definition of "direct cause" is implicitly acquired, according to the conditions on such a definition given below.

Whether Woodward's theory provides an implicit definition of causation is a model theoretic issue—and we currently have no definition of the class of models on which to interpret Woodward's theory. I will proceed below, first by outlining what I take to be a standard account of the basic conditions for an implicit definition, that I assume to be in line with Lewis (1970). I will discuss some properties of the class of models that could interpret T^W. I will apply the results first to the standard version of Woodward's theory, that relies on interventions such as they are defined in *Making Things Happen,* and then to an alternative, where we take interventions to be "soft." I will show for both of these alternatives that Woodward's theory don't satisfy the conditions for an implicit definition.

Let T^W be Woodward's theory, understood as above. I take it that T^W implicitly defines the relation "_ is a direct cause of _ relative to _"—which I will call C—if and only if the following two things hold. First, we take T^W to

be true of C *by stipulation*. This is the act of using T^W to define C. Second, the following must be true:

> **IMP:** For every model M of T^W, in the class $\mathscr{M}$ of models, $M \models \exists! R(T^W[R/C])$,

where R is a second-order three-place relation variable, and $T^W[R/C]$ is the statement acquired by substituting "R" for every occurrence of "C" in T^W. I will take the domain of R to be all sets of ordered triples of individuals from the domain of the theory. Thus, **IMP** says that, in each model of T^W, there must a *unique* three-place relation r such that $T^W[r/C]$ is true of it. Consequently, the implicit definition fails if in some model $M \in \mathscr{M}$ of T^W, there are two such relations. This amounts to a failure of the *uniqueness condition* on implicit definition.

8.4.4.2. *Applying Padoa's method to* T^W. Padoa's method is a way of proving that a theory fails to define a predicate, by failing the uniqueness condition (Craig 1956). Let $T^W(d, t_1, t_2, \ldots, t_n)$ be the theory, $d, t_1, t_2, \ldots, t_n$ the non-logical vocabulary of the theory, and d the definiendum. In our case, d is the causal relation C. We call $t_1, t_2, \ldots, t_n$ the (non-causal) *ground language* L of T^W. Let M^L be a model of the ground language, and M^{T^W} an *extension* of M^L to an interpretation of T^W. Padoa's method consists in showing that there exists a model M^L of the ground language that does not imply a *unique* extension M^{T^W}. In our context this means that, given some domain F of individual factors and an interpretation of the non-causal vocabulary of T^W on these factors, it is possible to extend the model with a relation C in two different ways. More informally, the non-causal facts relevant to the theory are then compatible, given T^W, with two different extensions for the causal relation C, and the theory thus fails to satisfy the uniqueness condition implied by **IMP**.

To apply Padoa's method to Woodward's theory, we need to know what its non-causal ground language L is. This requires some analysis of Woodward's conditions, since they are not stated formally. That is to say, identifying T^W's non-causal ground language is not a matter of identifying occurrences of the predicate "C" in the theory. Rather, to determine whether a certain condition in the theory is stated in the ground language or not, we have to consider what is required of a model that interprets this condition. I'll argue here that the only non-logical and non-causal part of T^W's vocabulary is the factor names "X," "Y," "Z," etc. Thus, a model M^L of the ground language of T^W consists in an interpretation of these terms over a set F of factors, and nothing more. It is then easy to show that the uniqueness condition is not satisfied by the theory.

We take T^W to denote the reduced formulation of $\mathbf{M_1}$, by which we mean that occurrences of "intervention" and "causal path" have been recursively eliminated throughout, in favor of their definitions in **IV** and $\mathbf{M_2}$. (We have also eliminated occurrences of "cause" in **IV** in favor of **CCS**.) Apart from the factor names "X," "Y," "Z," etc. and the relation C of direct causation, what remains of the non-logical vocabulary in T^W then occurs in the following conditions. First in *I2* from **IV**, that *there exists a value k for the variable I* such that X is disconnected from its other *causes* when $I = k$ occurs. This condition is

clearly causal. Second in *I4* from **IV**, that I *is statistically independent* of any other *causes* of Y. This condition, too, is clearly causal. Finally in $\mathbf{M_1}$, that *the probability distribution of* Y *changes* when $I = k$ occurs (and certain other things are held fixed). That this condition is also causal might require some explanation.

The difference that this change implies can (if we ignore the fixing interventions for simplicity) be stated as "$P_{I^{XY}=k}(Y) \neq P(Y)$," where P is the "default" joint probability distribution of the system (i.e., no interventions are taking place) and $P_{I^{XY}=k}$ is the joint probability distribution of the system when the intervention $I^{XY} = k$ on X with respect to Y takes place. Stated in this way, the condition explicitly refers to something (I^{XY}) that has been identified as an intervention, suggesting that the stated difference is intrinsically causal in nature. It's moreover a central result in causal inference theory that $P_{I^{XY}=k}$ can be derived from P together with certain causal information about the system (e.g., in the form of a causal diagram), and *not* in the absence of such information (Spirtes et al. 2000, sect. 3.7.2). Thus, any model capable of interpreting this condition must be a causal model, and "...Y changes..." is therefore a causal predicate, and thereby not part of L.

A model M^L, then, consists in a set F of factors and a relation N that interprets the factor names (i.e. variables) in L on this set. An extension M^{T^W} of M^L that interprets T^W contains in addition a set $\mathcal{P}$ of joint probability distributions over F, and a relation C over F. We can now construct two distinct models $M_1^{T^W}$ and $M_2^{T^W}$, that are both extensions of a model M^L, to show that the theory fails to implicitly define C.

Let M^L be an infinite set F of factors, named in L by variables "X_i," and let these factors have an order $(X_1, X_2, \ldots)$. Extend M^L to a model $M_1^{T^W}$ of T^W by adding a relation C_1 and a set $\mathcal{P}_1$ of joint probability distributions over F, in the following way. First let $(X_2, X_1, V) \in C_1$ and $V = \varnothing$. That is, X_2 is a direct cause of X_1 relative to V in this model. (Since there are no canceling paths in this model, it doesn't matter that V is empty.) T^W then implies that there is a further factor such that it satisfies the conditions on an intervention (variable) on X_2 with respect to X_1. Let X_3 be this factor, which means adding (X_3, X_2, V) to C_1. Next, T^W implies that there are joint probability distributions P and $P_{X_3=k}$ over F (where "$X_3 = k$" denotes an intervention on X_2 with respect to X_1) such that $P_{X_3=k}(X_1) \neq P(X_1)$. Add these to $\mathcal{P}_1$. Let X_4 be the possible intervention on X_3 with respect to X_2 then implied, and repeat the procedure indefinitely. In $M_1^{T^W}$, then, X_{i+1} is a cause of X_i, and X_{i+2} is an intervention on X_{i+1} with respect to X_i, for $1 \leqslant i < \infty$. (Since all factors lie on the same causal path in this model, each factor now satisfies the conditions for being an intervention on its causal successor with respect to that factor's causal successor.)

For our second alternative, extend M^L to a model $M_2^{T^W}$ by adding the empty causal relation $C_2 = \varnothing$ and let $\mathcal{P}_2$ be a set of any probability distributions whatever. As nothing is a cause of anything in $M_2^{T^W}$, and therefore nothing can be intervened on, T^W imposes no constraints on $\mathcal{P}_2$. Since $M_1^{T^W}$

and $M_2^{T^W}$ are different extensions of M^L, it follows that T^W doesn't implicitly define the relation of direct causation.

8.4.4.3. *What if we use "soft interventions"?* The fact that the change in the outcome variable that is referred to in T^W's necessary conditions for causation is itself intrinsically causal, and therefore can't be interpreted by M^L, apparently contributes to the ease by which we could construct a counterexample to implicit definition. The reason this change is causal, in turn, is that it is a *change under an intervention*, along with the fact that a standard intervention *modifies the causal structure* of the system intervened on. Specifically, the intervention on X severs X's connections to its default causes. This is expressed in condition *I2* from **IV**. As we saw in section 7.2.3, there is however also another type of intervention, which does *not* modify the causal structure of the system. These are often called "soft interventions," and it has been shown that a soft intervention on X with respect to Y suffices for the identification of a causal relation from X to Y (Eberhardt and Scheines 2007). A soft intervention satisfies all the condition in **IV** except *I2* (and they can be seen as a type of instrumental variable for X relative to Y).

Thus, if we eliminate condition *I2* from the theory—call the new theory T_S^W—and thereby regard our interventions as soft, the condition that the probability distribution of the outcome Y changes under the intervention event $I^{XY} = k$ can be replaced by the condition that the outcome Y is probabilistically dependent on the (soft) intervention variable I_S^{XY}, as determined by the *default* joint probability distribution P of the system. T_S^W still imposes all the other causal conditions on the causal relation, but "change" is no longer a causal notion in the sense it is when it denotes a change under a structure-breaking intervention. Thus, by opting for soft interventions rather than structure-breaking ones, thereby making "change" a non-causal notion, we have strengthened the constraints in terms of non-causal facts that the theory imposes on the causal relation. Woodward has moreover embraced the use of soft interventions in his theory (sect. 7.2.3). We therefore want to know if this strengthening T_S^W gives us an implicit definition of the relation of direct causation. Below, I will show that it doesn't.

To make the following discussion simpler to follow, I will ignore the possibility of canceling paths. Since, as explained above, I take it that the only purpose of *holding fixed* some factors in a set V is so as to eliminate any canceling of causal influences along different paths, this means that I will ignore also the relativization of the relation of direct causation to such a set V of factors.

T_S^W now implies that if X is a cause of Y, then there exists a factor Z that satisfies the conditions on a soft intervention on X with respect to Y. Let "$I_S^{XY}(Z)$" mean that Z satisfies these conditions. T_S^W also implies that Y probabilistically depends on Z, and we express this in the usual way. We can state this necessary condition for causation in T_S^W symbolically for conciseness:

$$
\begin{array}{ll}
 & \forall x, y(C(x,y) \rightarrow \exists z(\\
\textit{(i)} & I_S^{xy}(z) \\
 & \wedge \\
\textit{(ii)} & P(y|z) \neq P(y) \\
 &))
\end{array}
$$

Let M^{L_S} be a model of T_S^W's ground language. The condition indicated by *(ii)* above is not a causal condition, and must thus be interpretable by M^{L_S}. M^{L_S} then consists of a triple (F, N, P) where F is a set of factors and N interprets the variables on these factors, just as before, and P is a joint probability distribution over F. "$\neg\exists z(P(Y|z) \neq P(Y))$" may now hold or not hold relative to a model M^{L_S}. What this statement says is just that Y doesn't depend probabilistically on any other factor in M^{L_S}. If this statement holds in M^{L_S}, then it follows, in virtue of the necessary condition *(ii)*, above, that nothing is a cause of Y in an extension $M^{T_S^W}$ of this model. Non-causal facts can thus constrain the causal relation in the theory T_S^W in a way that they could not in T^W. (Remember, the corresponding notion of "change" in T^W was not interpretable in M^L, because the change refers specifically to a *causal* difference in the default and intervened-on versions of the system.) For example, if everything is probabilistically independent of everything else in M^{L_S}, then nothing is a cause of anything in any extension $M^{T_S^W}$. For this particular model, then, T_S^W determines a unique relation C, namely the empty relation.

We can now go on to show that, regardless of what F and P are in some model M^{L_S}, there is *always* an extension $M^{T_S^W}$ such that $C = \varnothing$. That is, the theory T_S^W is compatible with nothing being a cause of anything, regardless of what probabilistic dependencies hold between factors in P. Since some models are *also* compatible with some factors being causally related, we have then shown that the uniqueness condition is violated by T_S^W, too. These counterexamples are due to the fact that *(i)*, above, is an independent necessary condition for $C(X, Y)$ in T_S^W, that is moreover causal. Specifically, if there is no possible intervention on X with respect to Y, then X is not a cause of Y, regardless of any probabilistic dependencies. Whether there is such a possible intervention or not is in turn determined only in the extension of the model to T_S^W, by the causal relation C itself.

Take M^{L_S} to be a model with an infinite set F of factors, including X, Y, and Z, and a joint probability distribution P such that $P(Y|Z) \neq P(Y)$. Condition *(ii)* above is then satisfied for $C(X, Y)$. Now extend this model to T_S^W by adding a causal relation such that $C(Z, X)$, and such that there is no causal path going from Z to Y that doesn't pass through X, and such that Z is probabilistically independent of any factor that is a cause of Y by some path that doesn't pass through X. That is to say, add a causal relation such that Z is a soft intervention on X with respect to Y. Condition *(i)* for $C(X, Y)$ is thereby also satisfied. Thus, T^W implies that $C(X, Y)$, and we therefore add (X, Y) to C in $M_1^{T_S^W}$. T_S^W then also implies that there is a soft intervention Z' on Z with respect to X and a corresponding dependence $P(Z|Z') \neq P(Z)$ in M_S^L, and so on indefinitely, which we assume there is.

For our second extension $M_2^{T_S^W}$ of M^{L_S}, add the causal relation $C = \varnothing$. The situation is the same as in our treatment of T^W: since in this model nothing is a cause of anything, condition *(i)* is never satisfied, and condition *(ii)* therefore never goes into effect. No constraints are thus imposed on P, and $C = \varnothing$ is compatible with any P whatever, on the theory T_S^W. The two possible extensions of M^{L_S} shows that T_S^W also doesn't implicitly define C.

We have thus shown that even when we assume that interventions are soft, making more of the resulting theory's conditions interpretable on a non-causal model, the theory is compatible with *nothing being a cause of anything*, regardless of what probabilistic dependencies may hold. This result is stronger than just implying a failure of implicit definition: it shows that the theory isn't a plausible *approximation* of the meaning of causal claims, as might have been the case if counterexamples were restricted to exotic situations unlikely to occur in the real world.

Before discussing the root of this problem, I want to describe one more way of constructing a counterexample to implicit definition. I will use T_S^W again. This time I will start with a model of T_S^W and change this model in a way that leaves the underlying interpretation of the ground language intact. Consider once more the model $M_1^{T_S^W}$ in which $C(X, Y)$. In this model, at least one factor Z is such that $I_S^{XY}(Z)$ and $P(Y|Z) \neq P(Y)$. $I_S^{XY}(Z)$ further implies that $\neg C(Z, Y)$. I.e., if Z is a soft intervention on X with respect to Y, then Z is not a direct cause of Y. This is implied by *I3* from **IV**. Conversely then, any factor that *is* a direct cause of Y is *not* an intervention on X with respect to Y. Let now $M_3^{T_S^W}$ be a modification of $M_1^{T_S^W}$ according to the following. Add (z, Y) to C for every $z \in F$ such that $I_S^{XY}(z)$ in $M_1^{T_S^W}$, and remove (X, Y) from C. That is to say, in $M_3^{T_S^W}$ nothing is a soft intervention on X with respect to Y, because nothing that satisfies the other conditions on an intervention also satisfies *I3*. It then follows from T_S^W that X is not a cause of Y, which is also the case in $M_3^{T_S^W}$. $M_1^{T_S^W}$ and $M_3^{T_S^W}$ are moreover both extensions of the same model M^{L_S} of T_S^W's ground language, since F, N, and P are identical between them. This shows that for any model of T_S^W in which some factors are related as cause and effect, the theory is compatible with these factors not being causally related, on the same non-causal facts.

The method used to produce this last counterexample is essentially the method Woodward employs to account for the non-causal correlations between two particles that are in an entangled state according to the theory of Quantum Mechanics (Hausman and Woodward 1999). According to the standard interpretation of QM, a particle has no determinate state with respect to, say, its direction of spin, until the time when this property is measured. Moreover, the spin-states of entangled particles a and b are perfectly inversely correlated. Therefore, a measurement of the spin of particle a will "fix" also the value of particle b's spin. If we then consider the operation that measures spin-a as an intervention that *sets* this property in a (albeit to a value we cannot control), we might be led, on the interventionist theory, to conclude that spin-a is a cause of

spin-b. This conclusion would however be in contradiction with accepted physical theory—specifically with Special Relativity—because spin-a would then causally influence spin-b instantly regardless of the distance between the particles. It has been argued by some, therefore, that the entanglement cases imply a violation of some basic assumption about causal systems, either the Causal Markov Condition or faithfulness (Glymour 2006; Näger 2016). Woodward's solution is to instead interpret spin-a and spin-b, not as two distinct properties that may be causally related, but as in some sense a *single* "non-local" property of the system. This is, as I understand it, in line with the common understanding of this phenomenon in physics. That this property is non-local is a theoretical consequence—the two measurements certainly look on the face of it like measurements of two different things, at two different times and places. The implication is then, of course, that any direct cause of spin-a is also a direct cause of spin-b. Thus, nothing can be an intervention on spin-a with respect to spin-b, since condition *I3* cannot be satisfied, and therefore spin-a also cannot be a cause of spin-b (or *vice versa*). While this treatment appears to accord with mainstream physics, it nevertheless shows that, perfectly generally, whenever some X is a cause of some Y in a model of T_S^W, the theory is also compatible with X *not* being a cause of Y, given the very same non-causal facts, which is sufficient for a failure of implicit definition of the causal relation. To clarify, if "spin-a" and "spin-b" are variables, meaning names of factors, then they are clearly distinct. In Woodward's explanation of the entanglement case, they however refer to the same factor, and in virtue of this can't be independently intervened on, and therefore one can't be a cause of the other according to T_S^W. But this result isn't entailed by some constraint imposed by T_S^W, and the same may thus hold—as far as T_S^W is concerned—with respect to any pair of variables occurring in a causal theory.

8.4.4.4. *T_S^W and Reichenbach's Common Cause Principle.* The counterexamples to implicit definition presented in the preceding sections are an expression of the fact that T^W doesn't imply Reichenbach's Common Cause Principle (sect. 1.4.4). According to this principle, a probabilistic dependence between X and Y in P has *some* causal explanation: either X is a cause of Y, or the other way around, or they share a common cause. As we have seen, T^W is such that X and Y may be probabilistically dependent in P, but this dependence isn't due to any causal relations in the system at all. That this is a possibility is the reason why the theory is *always* compatible with X not being a cause of Y for any X and Y, and with nothing being a cause of anything at all.

In a paper preceding *Making Things Happen*, Woodward made the choice to reject the Common Cause Principle explicit:

> [I]t is possible for I to be correlated with some other cause Z of Y even though there is no causal connection between I and Z and even though Y and Z have no common cause. I thus reject what Cartwright [...] calls Reichenbach's principle according to which all correlations have causal explanations. (Woodward 1996, p. S30, footnote 3.)

While this statement doesn't occur in *Making Things Happen*, the condition *I4* in the definition **IV** of an intervention variable conforms to the quoted explanation. I.e., it is stated in terms of the statistical independence of the intervention on X from any variable that is connected to Y by a path that doesn't pass through X, and not in terms of the absence of a common cause of the intervention and Y. Woodward also points out that *if* we assume the Common Cause Principle, then condition *I4* reduces to the condition that the intervention variable and the outcome don't have a common cause (2003, p. 100). This suggests that Woodward's theory does not, and is not meant to, imply the Common Cause Principle. Perhaps the reason is to accommodate explanations such as that of apparent correlation that is due to entanglement, that we described above.

To somewhat complicate matters, the context in which this explanation of entanglement cases appeared was a defense of the universal validity of the Causal Markov Condition (Hausman and Woodward 1999). The CMC, in turn, implies the Common Cause Principle (sect. 6.3). The point there was, essentially, that on the proposed explanation of the entanglement cases, there *is* no correlation that violates CMC, since spin-a and spin-b don't denote distinct factors. Most writers seem to think that CMC doesn't hold for all causal systems (Strevens 2008, p. 190-191; Spirtes et al. 2000, p. 38; Arntzenius 1992). The argument in Hausman's and Woodward's paper is that, wherever CMC appears to be violated, this is a consequence of choosing the wrong referents for the variables.

> Apparent failures of the Markov Condition typically indicate limitations in background knowledge—that one is employing variables at the wrong level, or that one is failing to include relevant variables, or that one is treating variables or mechanisms as distinct when in fact they are not. (Hausman and Woodward 1999, p. 531.)

When appropriate things are taken to be causally related, then, CMC holds. In particular, the things named by the causally related variables must be properly *distinct*. This is a different notion of distinctness from the one I have assumed in this chapter, which consisted in logical, metaphysical, and mereological distinctness. It seems that the condition that properly causally distinct factors must satisfy, according to Woodward in this paper, is just that they can be *independently intervened on*. But the condition that for X to be a cause of Y, it must be possible to intervene on X with respect to Y, which implies that this is not also an intervention on Y, is already implied by T^W, so this particular condition on distinctness between factors introduces no new constraints on C.

In the next section I will introduce a further modification of Woodward's interventionist theory, such that this theory may succeed in interpreting the causal relation under certain assumptions, and I will discuss the problems with a proposal of this sort for Woodward's goals.

8.5. The price of success

That any X may fail to be a cause of any Y, regardless of what probabilistic dependencies hold between them, is a consequence of the fact that the existence of a possible intervention on X with respect to Y is a necessary condition for X to be a cause of Y, which in turn depends on the causal relation itself. This allows for violations of the Common Cause Principle. One way to possibly eliminate these counterexamples, then, is to introduce in the theory the rule that there exists a possible intervention on every factor with respect to every other factor. (This is condition **IC 2**, below.) If for any X and Y whatever, there exists a soft intervention I_S^{XY} on X with respect to Y, then what remains contingent in the interventionist conditions on causation is that $P(Y|I_S^{XY}) \neq P(Y)$. X is then a cause of Y if and only if Y is probabilistically dependent on I_S^{XY}. We can divide the new theory in two logical parts in a way that will be useful in the coming discussion. (Again I will ignore the possibility of canceling causal paths, for simplicity.)

IC 1: $\forall x, y \exists z (I_S^{xy}(z) \rightarrow (C(x, y) \leftrightarrow P(y|z) \neq P(y)))$

IC 2: $\forall x, y \exists z I_S^{xy}(z)$

I'll call the conjunction of **IC 1** and **IC 2** simply "**IC**." If **IC** interprets the causal relation, then for any set F of factors and joint probability distribution P over this set, **IC** is compatible with one and only one extension in F for the causal relation C. I have no proof to this effect, but will proceed in the discussion below under the assumption that it is the case. (If it is not the case, then my overall conclusions are not undermined.) We must note two things about **IC**.

First, for it to be possible to explain entanglement cases in the way Woodward does, there must be some constraints on what X and Y can be, i.e. on the causal relata. The constraint clearly cannot be that X and Y are causally distinct in virtue of being independently intervenable on, as this is already implied by the theory. In general, the constraints must likely be non-causal to do any work in the theory.

Second, we might ask at this point how **IC** could determine a unique causal relation, based only on the properties of a joint probability distribution? If **IC** does this, then it's because it supplies, in **IC 2**, a vast causal structure for free, as a matter of conceptual necessity. For every pair of factors X and Y, there is a cause Z of X that isn't a (direct) cause of, or share any causes with, Y. For Z and X in turn, the same thing holds of a fourth factor Z', and so on. This causal structure may be such that, given a probability distribution P, only one extension of C is compatible with the theory.

Here, I have ignored the qualifier "possible" when talking about interventions, which may appear at least misleading. These interventions are not presumed to actually occur, after all, only to be possibilities "in principle." However, as explained in section 8.3.2, we interpret possibility here in terms of the domain of possible values of the variables, not as applying to the factors,

denoted by the variables, themselves. It's hard to understand what the latter could even mean. Consider an example. We study some health condition in a sample consisting of a large number of people from different times and places. Some of these people are smokers, and we suspect a connection, so we introduce a variable "S" such that $S = 1$ for an individual if they smoke some minimum number of cigarettes or equivalent per day, and $S = 0$ otherwise. Some of the individuals studied may have lived in times or places where no one has heard of tobacco. They couldn't possibly have smoked, so necessarily $S = 0$, in some sense of "necessary," for these individuals. But S still has a value for them, and the variable S still denotes a causally relevant factor. Thus, to say that for some X and Y there exists a *merely possible* intervention I_S^{XY} is to say that it's not actually the case that $I_S^{XY} = 1$. It's *not* to say that I_S^{XY}'s presence in the causal structure is a mere possibility. I_S^{XY} must be part of the causal structure, for an intervention event $I_S^{XY} = 1$ to be possible at all. (Or perhaps better: the possibility, however remote, of the event $I_S^{XY} = 1$ introduces the factor I_S^{XY}.) Hence, what we get from **IC 2** is a vast causal structure—and not merely as a possibility.

Moreover, for this structure to determine (together with P) a *real* causal relation as a property of a type of physical system, the structure itself must be a property of that system, and not a feature of our beliefs, or our imagination, or of our formal theories. If it is the fact that we can *imagine* an intervention on X with respect to Y that determines (together with P) that X is a cause of Y, then this is not a realist theory of causation, but something closer to von Wright's conception of it (ch. 4). I will call the realist reading of **IC 2**, that I take to be required for **IC** to determine a real causal relation, the *ontological* interpretation of interventionism's necessary condition on causation. One of Woodward's goals is causal realism, and I will return to what is plausibly real in the interventionist account below. Now we need to discuss possible justifications for **IC 2**, and the ontological interpretation of it.

As already mentioned, the sufficient condition **IC 1** is uncontroversial under many different theories of what causation is, or of what causal claims mean, if "intervention" is understood in the proper way. Our concern is then the justification of **IC 2** and its ontological interpretation. First, it seems to me as though we can't *stipulate* that every causal system has this vast causal structure, as a matter of definition. This problem looks insurmountable on the face of it, from the philosophical perspective. Second, however, a necessary condition such as **IC 2** has, as we have seen, seemed to receive some support from the literature on scientific methods of causal inference, and I will focus on this situation here. Two sources in particular have appeared relevant in this respect. First, there are the causal modeling theories developed by Pearl and Spirtes et al., in which the effect of X on Y is *defined* as the change in the probability of Y that is a consequence of an intervention on X. It is moreover assumed that such a probability under intervention is defined for every pair X, Y in a model. Second, there are those, in particular within the Potential Outcomes framework of causal inference, who claim that there is no causation without manipulation or intervention. This looks like a claim close to Woodward's formulation of T^W,

in that it imposes possible intervention as a necessary condition on causes, and not, literally understood, on causal factors in general. But if we understand it as requiring a possible intervention on X with respect to Y for there to be a defined effect of X on Y, including a *zero* effect, then **IC 2** may seem to be implied, if we assume that the effect of any X on any Y, in any causal structure, is well-defined.

I think that, appearances to the contrary, there is no way of taking what is said about interventions in the context of causal models by Pearl and others, or in the context of the Potential Outcomes framework, by for example Miguel Hérnan, so that **IC 2** is implied.

As I noted in chapter 6, Pearl does not understand causation as conceptually or metaphysically dependent on interventions. He rather describes a causal relation from X to Y in terms of Y "listening" to X, in a way that suggests a primitivist notion of causation (Pearl and MacKenzie 2018). Nevertheless, he proposes an operational definition of "causal effect" in terms of intervention. I think that the right way to regard the "possible intervention" described in the context of causal models is as a type of *formal operation*, that is defined such that it can be applied to any variable in a mathematical causal model. The operation formally defines "X 's causal effect on Y" within the context of the theory, in that it extracts the causal influence of X on Y that is already encoded in a (sufficiently complete) model. Thus, that an intervention of this sort can be defined and applied to any factor modeled by a variable is a direct consequence of how these causal models are defined, and the assumptions made of them. And this is the *only* sense in which an intervention is always possible. They are possible formal operations on mathematical models, not possibilities that belong to the causal system itself. Specifically, a possible intervention on X of this kind is not a *cause* of X. I think that intervention variables, as they occur in Spirtes et al., should be understood in the same way (e.g., 2000, sect. 3.7.2). These variables are, in other words, technical devices. In contrast with the "regular" variables in the model, they don't generally denote properties of the causal system—i.e., *factors*. (This is one reason to insist on a distinction between variables and factors.)

The role of the superficially similar claim within the Potential Outcomes framework is quite different on closer inspection. (The status of interventions in these two theoretical contexts is a topic of current debate, see Hernán 2016 and Pearl 2018.) Miguel Hérnan argues that, for X to have a well-defined effect on Y, there must be a well-defined intervention on X—but Hérnan does not mean by "intervention" what Woodward (or Pearl) means by that term. It's central to Hérnan's argument that different interventions on the "same thing" can have different effects on the outcome of interest. Let body weight be the putative cause X, and the outcome Y death within some time period. Hérnan says:

> [I]f interested in the average causal effect of weight maintenance on death, empirical evidence suggests that some interventions would increase the risk (e.g., continuation of smoking), whereas others would decrease it (e.g., moderate exercise). (Hernán 2016, p. 677.)

What Hérnan describes is not possible if these interventions satisfy Woodward's conditions. Those conditions—condition *I3* specifically—are such as to ensure that any effect that the intervention has on death is mediated by body weight. The effect on death of losing weight by smoking, in Hérnan's description, clearly includes an effect of smoking on death that is not mediated by body weight. That is, Woodward's interventions are essentially surgical, while Hérnan's are not. The only way of modeling Hérnan's interventions in Woodward's framework, then, is to identify X, the putative cause itself, with the intervention. On this reading, Hérnan's claim that no causal effect exists in the absence of a well-defined intervention implies that, in his discussion, every *cause* is an intervention, or treatment (that is possible at least in principle). But Hérnan is also clear about the scope of this account. He is not claiming that these are all the causes there are.

> The potential outcomes approach was not designed to determine whether A is or is not a cause but to quantify the magnitude of the causal effect of A on Y. This quantification is only possible when the interventions are sufficiently well defined as argued above. In the absence of sufficiently well-defined interventions, the potential outcome approach is agnostic about causality. (Hernán 2016, p. 677.)

If what Hérnan calls an intervention is what is denoted by "X" in Woodward's theory, then we can understand Hérnan's claim about the importance of well-defined interventions as being about *the specificity of the cause* in any meaningful causal claim. That is to say, if we don't have in mind any particular intervention on X—in the sense of a particular way in which the value of X changes—then the cause in any claim about X's effect on Y will be too underspecified for the claim to have a determinate truth value. To insist on a well-defined intervention is, then, to insist that causes are sufficiently well specified for any effect to be well-defined, and to do this by using a heuristic device that is especially accessible to scientists familiar with experiments.

Moreover, some attempts at giving a more precise meaning to "well-defined intervention" in the Potential Outcomes context end up eliminating all reference to manipulation (and describing instead a covering-law account of causation):

> I would propose that a hypothetical intervention to set X to x is simply the specification, possibly contrary to fact, of the event or state $X = x$, such that we say that a hypothetical intervention is well-defined with respect to outcome Y, exposure X, setting $X = x$, and population P, if for each individual i in population P, there is a unique value $Y_x(i)$ (or distribution of values $Y_x(i)$ in the context of stochastic counterfactuals) such that the event or state $X = x$, along with the state of

> the universe and the laws of nature, jointly entail $Y = Y_x(i)$. (VanderWeele 2018, p. e24.)

The notion of an intervention here, then, appears to be employed as an aid to acquire sufficient specificity in causal claims. It seems right, therefore, to regard this intervention on X as a *precisification of* X. This precisification is, again, not a cause of X, but simply a description of X, and the necessity of giving this description in the form of a possible treatment is limited to a certain scientific context.

If an interventionist theory of causation succeeds in defining the relation of direct causation only when strengthened so as to imply that, for all factors X and Y in a model, there is another factor I that satisfies the conditions for being an intervention on X with respect to Y, meaning in particular that I is a cause of X, then this success depends on that "there is a possible intervention on every factor (with respect to every other factor)" is understood very differently compared to how I interpret Pearl's and Hérnan's use of "intervention," above. The ontological reading of "there exists a possible intervention on every factor," that is required for an interpretation of a realist causal relation, invites hard and currently unresolved questions about the senses of "exists" and "possible" here. This condition, with this reading, also finds no support in the literature on scientific methodology that we have considered.

8.6. What is real in the interventionist theory?

I mentioned above that, in recent writing, Woodward has focused on the methodological, or functional, aspect of his theory, and in a way that seems to downplay the semantic goals that he articulated in *Making Things Happen*. He has emphasized that thinking about causation in interventionist terms—i.e., in terms of what would happen in a hypothetical experiment—has methodological benefits. He gives examples of what this can help with:

> (i) Pick out the target information we are trying to discover when we engage in causal inquiry (the outcome of a hypothetical experiment) and in doing this also help us to clarify the original causal claim or make it more precise.
> (ii) Show that certain causal questions we may be tempted to propose are not answerable, at least with available data—not answerable either because they do not correspond to any possible experiment or because the actually available data cannot provide answers to questions about what outcome of the hypothetical experiment would be.
> (iii) Clarify and evaluate some of the methods used to infer to causal conclusions, particularly in the case of non-experimental data. Very roughly, the idea is that we ask whether the data are such that (in conjunction with appropriate other assumptions) they can be used to infer what the results of the associated hypothetical experiment would be if we were to perform the experiment, although in fact we don't or can't actually perform the experiment. (Woodward 2015b, p. 3587.)

The reason for insisting on an interventionist frame of mind, then, here looks very similar to the reasons I attributed to Miguel Hernán, when he insists on well-defined interventions (and Woodward appeals to a similar-sounding scientific source). As I mentioned above, I don't think that these reasons support the ontological reading of the claim that there is a possible intervention on every factor, that in turn has such implications for real causal structures as to eliminate our counterexamples to an implicit definition. What I think that these reasons do support, is a reading in which "there is a possible intervention on every X" means 'there is a formal operation applicable to every variable "X" in a causal model' or "there is a precisification of every factor X in a theory, when there is a consistent effect," neither of which are *causes* of X. On these readings, the interventionist theory of causation does not single out a unique causal relation.

Causal realism seems to commit us to the existence of a causal relation in the real world, independently of whether interventionism defines this relation or not. Woodward defends causal realism in his criticism of Menzies's and Price's manipulationist account. He there says that

> it is facts about how the world is and not facts about my expectations or projective activities that determine what will happen to my longevity in the experiment in which I purchase life insurance. (Woodward 2003, p. 119.)

And then:

> Contrary to what many philosophers have supposed, a commitment to some version of realism about causation (in the sense that relationships of counterfactual dependency concerning what will happen under interventions are mind-independent) seems to be built into any plausible version of a manipulability theory. (Woodward 2003, p. 120.)

Let's assume that agreeing with Woodward here commits one to the existence of a real causal relation C_R, with a determinate extension. Then, given Woodward's definition of an intervention, **IC 1** looks like a plausible constraint on C_R. That is to say, **IC 1** is plausibly *true of causation in the world*. This is in fact necessary, if the theory is to explain the epistemic utility of causal experiments. C_R determines univocally, of course, what is a cause of what in the real world, but also, due to **IC 1** being true of it, what is an intervention on what and what any causal consequences of interventions are. C_R determines also the truth values of every meaningful causal counterfactual conditional, as is made explicit in some theories of causal inference. Hypothetical causal reasoning then requires a corresponding hypothetical causal relation C_H. In contrast, if C_R violates interventionism's *necessary* condition on causation, i.e. the condition that states that there is a possible intervention on every cause or on every factor—I'll call this condition **NEC**—then this doesn't seem to have any adverse consequences for our causal realism, our understanding of experiments, or the interpretation of causal counterfactuals. If there is some defensible version of **NEC** at all, then—as I argued in the previous section—I think it is as a statement about formal operations, or specific descriptions

of causes, not as a statement about causal structures. C_R and **IC 1** is then what is *real* in the interventionist account. And this seems like an important difference in the logical parts of the interventionist proposal.

A theory that fixes on what is real and well-supported in the interventionist proposal could then retain **IC 1**, viewed as a constraint on any causal relation C, that itself is taken as a theoretical primitive. **NEC** is rather taken, if at all, in what I think is Pearl's vein, as a statement about the existence of a certain formal operation, that can be applied to any variable that occurs in a sufficiently specific causal model, to extract the effect of that variable on some other variable, that is encoded in the model. These "formal interventions" are then different in essence from the interventions mentioned in **IC 1** and **IC 2**. I'll label this kind of interventionist theory "**ICI**."

ICI can't be conceived of as a theory of causation, or of what causal claims mean, because it has no necessary condition for causation, and a causal relation is assumed as a theoretical primitive. It in particular doesn't imply the necessary condition for causation, that I took in chapter 1 to characterize manipulationist and interventionist theories of causation, so it's specifically not *that* sort of theory of causation. What **ICI** contributes is an understanding of *interventions*, given certain conditions on a causal system such as CMC, faithfulness, and modularity, that may hold only contingently. Interventions are events that isolate the effect of one factor on another. The theory states the causal conditions under which this is the case, in Woodward's definition **IV**. Causal experiments, and manipulations in general, succeed to the extent that they satisfy these conditions. I think, therefore, of **ICI**—the real part of interventionism—as *a theory for causal inference under intervention*, or just a theory of intervention.

I think that **ICI** substantially enhances our understanding of manipulation and experiment, and in that particular sense it tells us something important also about causation. In the next section I will apply **ICI** to a question from the epistemology of causation, in a way that aims also to alleviate certain worries about regarding causation as a primitive.

8.7. Manipulation and our acquaintance with causation

8.7.1. The revenge of the circles. If we assume as a primitive a relation C of direct causation, and reject (at least the reality of) the necessary condition **NEC** (which, to repeat, states that there is a possible intervention on every cause or on every factor), then we no longer have circularity in our definitions of "intervention, " "causal path," and "contributing cause." But this doesn't resolve the methodological and epistemic regress problems in the interventionist account detailed by Baumgartner (2009). That is to say, there are epistemological issues that **ICI** can't—and, I think, shouldn't be expected to—address. **ICI** can only generate new causal knowledge when we know that some event is an intervention, which is knowledge about the causal relation. Can we, then, ever know that something is an intervention? Woodward suggests at times that randomized controlled trials (RCTs) have this role. He describes

> one of the main virtues of a randomized experiment (which the notion of an intervention is meant to capture): when you successfully carry out such an experiment you remove correlations between the putative cause and effect that are due to all potential confounding causes, even those that are unknown or unobserved or "invisible" [...] (Woodward 2008, p. 203.)

But, as Baumgartner also observes, this alone doesn't suffice as an explanation of how we can *know* that something is an intervention, under the classic expectations on knowledge. It all comes down to whether the experiment in question really was an RCT, and knowing this means knowing that the selection for treatment was independent of every cause of the outcome, that may otherwise confound our results. (We have often failed to know this.) This is intrinsically causal knowledge. The problem appears to mirror the general one, that we can't infer causal information without having some causal information in our premises, which suggests an infinite regress. Thus, what we need is a more basic epistemic story, that bypasses causal inference as well as explicit justification altogether. I think there is a story like this to be had, and I think it can be substantially informed by our theory of interventions. (But it's not the old story of free and voluntary actions, that are globally causally exogenous, and therefore constitute will-directed, uncaused injections of causal influence into the natural world.)

Those philosophers who subscribe to a "singularist" view of causation, as opposed to the traditional Humean regularist view, have often been open to the idea that causation can be (veridically) *experienced* on a single occasion (Ducasse 1926; Armstrong 1962; Anscombe 1971). The focus in these arguments has most often been on a passive experience of causation, in vision or some other sensation (such as touch). The philosophy that discusses agency and causation, on the other hand, has mostly been concerned with causal relations between volition and action (e.g.: Hume 1888, Appendix: "To Be Inserted In Book I. Page 161. Line 12"; Chisholm 1966; Mumford and Anjum 2011). The goal of the sort of account I will sketch here is different from both of these traditions. The goal is rather to show the possibility of theory-independent and non-inferential experiential knowledge of causal relations between external (to the agent) events, that is generated by our own manipulations. What motivates such an account is the search for a source of some initial causal knowledge, that can then be further employed in explicit causal judgments and inferences. It thereby also seems to relate importantly to our understanding of successful causal experiments, in science as well as in everyday contexts. The need for some account like this is explicitly recognized by Woodward. In relation to the fact that any reliable causal judgment must rely on prior knowledge of causal mechanisms, he says:

> [A]t some point, subjects must learn the mechanism information to which [...] they are appealing when they make actual cause judgments. Even if they learn it as a result of being told by others, at some point someone in this chain of communication must have learned it from other sources, which

> presumably include experiences of various sorts. (Woodward 2018, p. 134-135.)

Cartwright, too, invites the possibility of experiential knowledge of causation, perhaps for related reasons.

> Is there a reasonable source in experience for our concept of single case causation? Again, Professor Anscombe thinks there is. But this is not an issue I will discuss here, except to hope that she is right. (Cartwright 2000, p. 49.)

8.7.2. How manipulations may satisfy the conditions on interventions. Since it contains much empirical speculation, including interpretation of results from cognitive science, I take what follows here ultimately as a "proof of concept," of the application of our theory of intervention to the issue just discussed. What I do commit to is that something along these lines is *possible*, in a strong and interesting sense. Relative to the traditionally dominant view of causation in analytic philosophy, this looks like a bold claim in and of itself.

Our starting point then, is causal realism, and our insights into the conditions under which an event is such that we can infer causation from a correlation (i.e., the event is an intervention). I will also think of agents in a completely naturalistic way, as a class of causal systems, with certain characteristic physical and causal features. The argument is then going to be that, in virtue of these characteristic features, and under a certain epistemology of sense-based knowledge, our *sense of agency* can be a reliable source of non-inferential causal knowledge.

A manipulation, we have said above, is an intentional action essentially involving a bodily movement, with some direct result in the external world. Under our naturalistic, causal view of agents, I will assume that what is meant by an action being intentional is just that it results from the operation of the agent's conscious decision mechanism (which is likely some part of the brain). This will stand in the place of "volition." The current proposal will, moreover, only make sense if we limit the direct result of an action in this context to the thing pushed or pulled or held in place by the associated bodily movement. It is thus a "physically direct" result. These are button pushes and the like, rather than, say, selections for treatment—but even selections for treatment must begin with some body part colliding with some external object, whether it is a mouse button when running a computer program, or a paper note when pulling names from a hat. (Or, I suppose, just the atmosphere, when giving a verbal command.) What's important here is that no more remote effects of actions count as direct results. They will rather be "outcomes."

This is then the event that should, sometimes, satisfy the conditions on an intervention, as they are given in **ICI**: a bodily movement, that is an effect of a decision by the agent. Note that we are not aiming for infallibility, but reliability. We state the conditions on an intervention again, informally.

> **Int:** A manipulation M is an intervention on its physically direct result C with respect to the outcome E *if and only if*

(1) M is a cause of C.

(2) M is not a cause of E unless by way of C.
(3) M does not have a common cause with E.

Here I ignore condition *I2* in **IV**, which says that M is the *only* cause of C, since we know that a "soft" intervention, that doesn't disconnect C from its pre-manipulation causes, is sufficient for our result. Since a manipulation must begin with a physical interaction, it seems natural to think that 1 is satisfied if the bodily movement introduces a force on the external object that takes it out of its current trajectory (which might consist in it just sitting there, unmoving). If this force is moreover precise, in the sense of narrowly directed and of an appropriate magnitude, then it is also likely that 2 is satisfied. It is then "surgical" in a particularly straightforward sense. 3 is more complicated, and this is where the physical constitution of agents and their decision mechanisms must, I think, play a part. Intuition tells us that if we freely (as we experience it) manipulate some factor C, and some more remote factor E systematically follows our movements, then it is unlikely that this is due to a common cause of E and the manipulation M. On a time-honored account of free action, one might want to explain this by appealing to global exogeneity: our action is free in the sense of having no natural causes at all, thus it has no common cause with E. But we know that a more restricted type of exogeneity is sufficient, that is moreover compatible with our naturalistic view of agents. What is required, then, is an explanation of why it is unlikely specifically that the manipulation has a common cause with E.

An intuitive idea is this: if some nearby external event caused me to perform this action just in the way that I am, then I would know it. It would have to have been on purpose (as in the button-horn example, described in section 8.2.2). We know that this isn't necessarily true—we can be subject to external influences without noticing—but it may still hold true *most* of the time. Here, then, are two physical features that I take to be essential to agents. (1) Their decision mechanisms are not open to arbitrary causal influences from the environment, they are *causally isolated* to a significant degree. (2) External causal influences that *do* reach the decision mechanism, are to a large extent fed as *data* to it. In the case of humans, our skull isolates the decision mechanism. Some external causal influences penetrate this barrier in perfectly normal situations, but it happens mainly at the openings where our eyes and ears and other sense organs are connected. I take this sort of configuration to be an essential aspect of agents. If its decision mechanism is largely open to arbitrary causal influences from the environment, that are not accounted for in a systematic way, then it probably won't make good decisions. If we accept this very general story so far, and its relevance to the problem at hand, then we have reason to believe that, unless the agent decides to cue their action to some external event, it is unlikely that the manipulation and the correlated remote event have a common cause in the immediate surroundings. I assume that "unlikely" is good enough for our purposes.

But this only takes care of possible *nearby* common causes. We have historically tended to think of chains or paths of causation that stretch far back in time. Given that any agent has a limited ability to perceive and remember,

what's to say that the manipulation and the outcome don't have a common cause somewhere in the remote past? Go back far enough, and you may find some cause of the manipulation event taking place before the agent even existed. Intuition instantly balks, I think, at the idea that a strong correlation between a manipulation and a nearby event is explained by a remote common cause, but the question is why.

I believe that a precise answer requires a *quantitative* notion of causation, something that a theory for causal inference provides. This since, on my proposal here, we must be able to speak of causal influences "petering out" over time, due to an accumulation along the way of independent causal influences on the outcome. This is straightforwardly captured in a model of additive *causal noise*. And causal equations can be such models. In the causal equation "$Y \stackrel{c}{=} f(X,U)$," we have interpreted the noise variable "U" as a random variable that encodes the exogenous causal influences on Y, that are independent of X. We can understand the idea that the influence of a causal factor X on a type of outcome peters out over time as the fraction of the variance in the outcome Y that is caused by X (rather than by Y's exogenous causes U) being a strictly decreasing function of the time between the X-event and the Y-event in this type of system. In a straightforward linear case, modeled by "$Y \stackrel{c}{=} b_{XY}X + U_Y$," we might use as a measure the coefficient of determination R^2, where $R^2 = 1 - \frac{Var(U_Y)}{Var(Y)}$. This is sometimes interpreted as the fraction of the variance in the outcome Y that is explained by the variance in X. Under our causal interpretation of the association between X and Y, it seems natural to understand it as the fraction of the variance in Y that is *caused* by X. (For non-linear and non-parametric cases, some other measure is needed.) To show that X's influence on Y peters out as a function of the time between the X-event and the Y-event, relative to a type of system, we need two general assumptions:

(1) Every factor on a causal path is affected by some exogenous causes, that are independent of the exogenous influences on the other factors on that path.
(2) Temporally extending a causal path introduces new intermediate factors on it.

1 is a standard assumption in causal modeling. 2 goes well with a physical understanding of the continuous propagation of causal influence through spacetime. We now want to compare R^2 for Y before and after we extend the path between X and Y. Let $Y \stackrel{c}{=} b_{XY}X + U_Y$ in this type of system, as we said. We take extending the path between X and Y to entail the addition of an intermediate factor Z right before X on the path, by assumption 2. To keep the system otherwise unchanged we take to mean keeping Y's noise variable U_Y and its coefficient b_{XY} unchanged. Let $Z \stackrel{c}{=} b_{XZ}X + U_Z$. From Z being a causal intermediary between X and Y, it then follows that $Y_{Ext} \stackrel{c}{=} b_{XY}b_{XZ}X + b_{XY}U_Z + U_Y$ in the extended system. We need to make one more assumption at this point, namely that Z is not a "signal booster." A signal booster amplifies the causal influence in the path. Such things clearly exist (for example in electronic communication systems), but they are presumably unusual in the natural world.

We can represent this no-signal-booster assumption by "$|b_{XZ}| \leqslant 1$." (That the model by design only includes causes on the right hand side entails that $b \neq 0$ for every coefficient b.) Since moreover $Var(U_Z) > 0$ by assumption 1, it follows that $1 - \frac{Var(U_Y)}{Var(Y)} > 1 - \frac{Var(b_{XY}U_Z + U_Y)}{Var(Y_{Ext})}$. That is to say, less of the variance in the outcome is caused by X in the extended system, and thus the correlation between X and Y is weaker.

It then follows under these assumptions that the strength of the correlation between X and Y will diminish with time in this kind of system, and therefore that a high degree of correlation is improbable if the distance between X and Y is sufficiently great, or if noise accumulates sufficiently fast. Crucially, under the assumption that the exogenous causal influences on the factors on *different* paths are jointly independent, the same holds of the correlation between Y and another factor R, that is located *near* Y, when Y and R are only connected by having X as a common cause. That is, the strength of the correlation between Y and R is then a strictly decreasing function of the distance to their common cause.

If we know something about this function, then additive causal noise can be epistemically useful. Consider an ordinary signal wire. Such a wire has some noise characteristic that may be well known. Thus, if we can measure the strength of the signal at the source and at the receiver, then we can estimate the length of the wire. Likewise for the total length of wire connecting two receivers to a common source. The wires are metaphorical—any realistic environment is causally noisy, and any effect of some event will, under our assumptions, be drowned out by noise eventually.

For this to be useful for our purposes the noise function—the rate at which the effect will diminish relative to independent exogenous influences—must be sufficiently *fast* in a stable manner. Comparison must be possible between different occasions. This condition seems *prima facie* unlikely to be satisfied. Surely, the longevity of the "signal" depends on the size and robustness of the events in the system, and the causal "busyness" of the surroundings? To illustrate, we may be able to predict an effect of a cosmic event that is a hundred thousand years into its future, but lose track of the effects of a particular chemical interaction within seconds. But there is an environment that is always around when manipulations take place, and that may be sufficiently noisy between occasions to be useful, and that is the brain. If this is right, then the existence of a very remote common cause may be an unlikely explanation for a significant correlation between our manipulation and an external event, due to the complexity of our minds (brains), where this vast causal complexity can be understood in terms of noise—noise just being causal influences on the action that are independent of that remote external cause of it. What our intuition latches onto, in other words, is the implicit belief that our manipulation wouldn't correlate so strongly with that external event, if they only shared some remote common cause (that we aren't aware of).

To summarize: it is unlikely that the manipulation has a nearby common cause with the correlated external event, because most of the time the agent would know if this was the case. And it is unlikely that the manipulation and

the correlated external event have a common cause that lies outside of the agent's sphere of attention, because it is unlikely that such a remote common cause would give rise to such a strong correlation between the manipulation and the external event. I propose that this idea could be made mathematically precise in a model that treats the accumulated causal noise U in an equation $Y \stackrel{c}{=} f(X, U)$, stating the external cause X's effect on the action Y, as a strictly increasing function of the time between X and Y, where this noise function is sufficiently fast relative to a certain system that includes the agent's decision mechanism.

Thus, we have seen some reasons to believe that, when there is a strong correlation between a precise manipulation of X and some external factor Y, and the agent is paying attention and not intentionally correlating their action with some nearby external event, then it is likely that the manipulation satisfies the conditions on an intervention on X with respect to Y. However, even if this happens to be right, it doesn't so far amount to any new *knowledge* on the part of the agent.

8.7.3. Sensing causation. Clearly, the suggestion here isn't that the argument I just gave, in defense of the idea that manipulations are likely to be interventions under certain seemingly unremarkable circumstances, is any part of human causal *reasoning*. I doubt that it's recognizable as such, and anyway, we are after *non-inferential* causal knowledge. This story isn't about *reasons* we might have for any causal beliefs or judgments, but about the possible evolution of a sense for causation. The sense I mainly have in mind is usually called the *sense of agency*. But I think there are several senses that are essentially involved in our experience of causation. Before we go on, we need to say something general about sense-based knowledge in this context.

First, the key to ending the regresses of inference and justification is, it seems to me, a causal and reliabilist understanding of knowledge. So, we adopt the following theory of sense-based knowledge:

> **K:** A has direct, experiential knowledge of the presence of O based on a sense S *iff*
>
> (1) O is present,
> (2) A believes that O is present, and
> (3) A's belief that O is present is caused by a sense S, such that sufficiently often when this belief is induced in the agent by S, O is present. (I.e., S is reliable.)

The sense does not need to be infallible, of course, and—crucially—the agent doesn't need to be able to *understand* or *justify* the reliability of the sense, to be the object of true knowledge ascriptions. In particular, this allows small children, animals, and perhaps artificial agents, to acquire direct, experiential knowledge.

Second, if we have a sense for causation, then it isn't like the so-called five senses. Elisabeth Anscombe—who famously claimed that we can observe particular instances of causation—nevertheless conceded one of Hume's points.

> Hume presumably wants us to 'produce an instance' in which [causal] *efficacy* is related to sensation as *red* is. It is true that we can't do that; it is not *so* related to sensation. (Anscombe 1971)

I take the point here to be that a causal relation (or its necessity or efficacy) is not presented before us quite in the way redness is, when we look at a red object under normal conditions. If we do have a sense for causation, then it is plausibly a complex, higher order sense. We can perhaps compare to the case of three-dimensional space. We have a sense for this: our *depth perception*. This is a complex sense, that is triggered by a multitude of diverse cues, such as:

- stereoscopic vision
- movement in the visual field when we move
- level of saturation of colors of objects
- focusing of the eye
- differences in sounds and echoes

(For more—much more—see Howard 2012.) Despite its complexity and reliance on more basic senses, such as vision and hearing, no one seems to think that our depth perception is generally unreliable, or that what it reports isn't real. In particular, the fact that we can't *derive* three-dimensional space, or *prove* its existence, from the cues that induce beliefs about spatial distance, doesn't seem to make us question its reality. A sense for causation, I think, would have to be a sense of this kind, that depends on multiple cues, provided by more basic senses, where correlation plays one central part, and where the regular veracity of the beliefs that it induces is sufficiently supported by abductive reasons.

We may now return to our conditions on interventions. I have argued that it's reasonable to think that manipulations may regularly satisfy these conditions, when the manipulation is precise and the correlation between the manipulation and a nearby external event is strong. Do we have any *sense experiences* that match up with these conditions in an appropriate way?

I think that the senses of *touch* and *proprioception* can vouch for the first two conditions: that the agent caused the putative direct result with their manipulation, and didn't cause much else by that action (and thus probably not the outcome, unless by way of the direct result). We said that these conditions are likely to be satisfied when a bodily movement exerts a precisely directed and measured force on an external object, that affects its trajectory. Touch and proprioception are the senses by which we perceive forces acting on our bodies.

> Touch is the sense by which pressure exerted on the skin is perceived, pressure being a function of force, namely, force divided by surface area. Proprioception (or kinesthesia) is the sense through which we perceive the position and movement

> of our body, including our sense of equilibrium and balance, senses that depend on the notion of force. (Wolff and Shepard 2013, p. 174.)

Wolff and Shepard cite research showing that humans are very good at estimating the magnitudes and directions of forces applied to their hands, supporting the claim that we can have a good idea about whether the surgicalness condition on interventions is satisfied. (Unsurprisingly, seeing how actual surgeons are human beings.)

The following example illustrates how having these senses put an actor in a different epistemic position, vis-à-vis causation, as compared to a passive observer. Imagine that you are driving a car and that I am your passenger. I can see your hands and the car's steering wheel move back and forth in unison. However, I can't know based on this observation if it is you who are moving the steering wheel with your hands, or if the wheel is moving by some other cause, and moving your hands along with it. (Maybe we are in a self-driving car.) *You* can know, though, because you feel a force acting on your hands and arms. If the direction of that force is the opposite of the direction that the wheel is moving, then it is you who are moving the wheel. If the force has the same direction as the wheel, then the wheel is moving your hands.

That touch is involved in our experience of causation was famously suggested by D. M. Armstrong (1962). Proprioception has been defended as a proper sense by Brian O' Shaughnessey (1995) and Ellen Fridland (2011). Mumford and Anjum discuss its role in causal perception at length, but their primary target is the relation between volition and action (2011). Our current target is apparently more philosophically modest: just that these senses can tell us that our bodily movement caused a certain nearby change in the external environment. But this is no small thing when you think about it. Because, it seems that knowing that you had to apply a certain amount of force to move something (or keep it still), entails knowing that it wouldn't have moved (or remained still), had you not applied that force. It's then something close to directly experiencing the counterfactual implication that the causal claim entails.

Our larger purpose is establishing the plausibility of experiential knowledge of causation between external events, and knowing that our manipulation caused its direct result (and not much else) is a means to that end. The experience that provides for our overall goal is, I think, our *sense of agency*.

In a 2012 article, the cognitive neuroscientists Patrick Haggard and Valerian Chambon explain that "[t]he term 'sense of agency' refers to the experience of controlling one's own actions, and, through them, events in the outside world" (Haggard and Chambon 2012, p. R390). The sense of agency is mainly associated with complex, intuitive tasks, such as driving a car or playing the piano. There are also clarifying contrast cases in the context of cognitive pathology, such as when someone has *too much* of the sense of agency in certain cases of schizophrenia, or *too little* of it for example in cases of "alien hand syndrome."

As to the this being a proper sense, and not after all just another case of causal inference, studies in cognitive science have identified what is called an "intentional binding effect," that has been taken as a cognitive side-effect of the presence of an implicit sense of agency. This effect consists in a distortion of the time, estimated by a person, between two events, when one is the person's action and the other an event over which that person has a sense of control, by way of that action. By using this measure, and comparing it with subjects' explicit self-reports of agency in a separate trial, Nao Saito and collaborators reported in a 2015 paper that "by comparing the two distinct methods of measuring the sense of agency, we found supporting evidence for the dissociation of the explicit judgment of agency from the lower-level experience of the feeling of agency." (Saito et al. 2015, p. 6).

I take this as supporting the general claim that we do have an implicit sensation of agency—of our actions being the causes of certain events—that is something other than a *judgment,* based on a causal inference.

While much of the research into the nature of the sense of agency concerns the sense of control over ones own bodily movements, the sense extends, as we have said, to external events. Since my interest is in this sense as a possible source of *new* causal knowledge, we need to assume that it extends to a certain restricted class of external events, over which we may then have a sensation of "I made that happen," or of "authorship," as it is sometimes called. It is those situations in which we don't have strong beliefs in advance about what will in fact be the effect or our action. My favorite example is one where we figure out the controls for a computer game—as in, "what effect does pressing *this* button have?" I don't know of any research relating to how the sense of agency actually performs in these kinds of cases specifically, but will proceed under the assumption that what is said for the general cases is applicable.

My claim goes beyond saying that something in our cognitive system makes us believe that our manipulation caused a remote event. I'm claiming that this is a *reliable* process—at least to an epistemically useful degree. The argument must then be that, among the likely diverse range of triggers for the sense of agency, there are important ones that tend to co-occur with situations in which our manipulation satisfies the conditions for being an intervention on its direct result, with respect to the correlated external event. I argued above that *strong correlation*, between the manipulation and the external event, together with a precisely directed and well-measured application of force on the direct result of the manipulation, is sufficient for it to be probable that the manipulation is an intervention. It is also more than likely that strong correlation is one of the primary triggers of the sense of agency. From a biological point of view, if a strong correlation between a manipulation and an external event tends, sufficiently often, to be present only when this correlation is *not* due to a common cause, and if knowing some remote effects of our actions increases fitness, then this is something that could be selected for, in the evolution of a sense for causation. I don't think that such a sense is going to be *very* reliable. In particular, it may be cheaper for the organism to deal with frequent false positives, than to evolve a very precise instrument for identifying remote effects

of actions. What getting it right "sufficiently often," and the sense of agency being "sufficiently reliable," mean exactly, would be determined by survival and procreation rates.

Regardless of the scientific plausibility of this particular account of non-inferential causal knowledge, I take it to show that something like this is possible, in the sense of "physically" and even "biologically" possible. If it, or something like it, is true, then it provides us with a fallible but nevertheless useful source of some simple causal facts. If we moreover take these to be facts about causal relations between particular events, in virtue of their types (captured by what factors they involve) and their causal context (i.e., the causal system in which they occur), then these facts can be employed in subsequent causal inferences. (How the story continues, after we have established the presence of that first causal arrow, I won't speculate about here.)

As a final note—if we have direct experiences of causation, then this would seem to make "causation" an empirically legitimate concept in the absence of any definitions, reductive or otherwise. It might, that is to say, make the idea of taking C_R as a primitive, in the way suggested above, more appetizing to some.

8.8. Summary of conclusions

I have argued in this chapter that the central problem of the older manipulationist theories of causation, up to and including that of Menzies and Price, is that they fail to provide a sufficient condition for causation, by not excluding some possible confounders of a correlation between the manipulation and some remote event.

In the modern incarnation of manipulationism, developed mainly by James Woodward, sufficient conditions for causation *are* in one sense provided. But partly for this reason, the theory is ineliminably circular in its explicit definition of causation, and the conditions imposed by the theory are moreover too weak to constitute an implicit definition. The theory therefore cannot provide the meaning of causal claims in the way that Woodward means to in *Making Things Happen*, i.e., by providing a definition of the relation of direct causation.

I've argued that interventionism's characteristic implication, which is that there is a possible intervention on every cause, must be strengthened to stipulate a possible intervention on every factor with respect to every other factor, for implicit definition to be possible (without claiming that this is sufficient for success), and that this condition must also be given an ontological interpretation, in the sense that the intervention factors (that provide for the possible intervention events) are a proper part of the causal structure. But this ontological interpretation of the interventionist necessary condition on causation, that I've called **NEC**, receives no support from what I think is a plausible reading of superficially similar claims about interventions in causal inference theory. I know of no other justification for **NEC**, which I take to be *prima facie* implausible.

However, **NEC** can be subtracted from the theory. The remainder—which is the theory I call **ICI**—has, if combined with the primitive assumption of

some causal relation C, the character of a theory for causal inference under intervention. It therefore can also fruitfully be regarded as a *theory of intervention*. As **NEC** is the implication that I've taken as defining a theory as a manipulationist or interventionist theory of causation, **ICI** is specifically not such a theory. **ICI** must be held to be true of causation as it is in the world, at least when certain conditions such as the Causal Markov Condition and modularity hold, for the theory to support sound causal inferences. If we insist on retaining **NEC** on, say, pragmatic grounds, we must nevertheless recognize that it doesn't have this status, of stating something *real* about causation. This is a crucial difference between interventionism's sufficient and necessary conditions, in my mind, with direct implications for the interpretation of the theory.

Those are, then, my conclusions as regards the role of manipulations and interventions in this class of theories of causation. As to the true role of manipulation in our relationship to causation, I have argued that the theory **ICI** can substantially contribute to our understanding of it. I've argued that this theory can help explain how we acquire causal knowledge, not only through inferences, but through direct experience, because it provides the conditions under which a correlation under a manipulation implies causation—conditions that I think match up with certain sense experiences in a sufficiently reliable way to be epistemically useful. Thus, I take myself to defend Woodward's claims that, "it seems plausible that many voluntary actions do, as a matter of empirical fact, satisfy the conditions for an intervention," and "subjects who are guided by [roughly, their sense of agency] will make fairly reliable causal inferences" (Woodward 2007, p. 22, 30.)

So, while I reject manipulationism *and* interventionism, viewed as competitors on the stage of philosophical theories of causation, I strongly agree with their tradition of giving manipulations a unique role in our acquisition of causal knowledge, and our formation of causal concepts. If this is, somehow, what some manipulationists have had in mind all along, then we are in agreement—but this is an *epistemic* role, nothing more.

Nevertheless, this leads to a view of the role of manipulation, and by extension experiments, that differs significantly from that in traditional, regularity-oriented analytic philosophy. The view I have in mind is described by Edwin McCann in his review of von Wright's theory.

> [I]t is not the fact of the experimenter's interfering with the course of events that is important to the epistemic status of an experiment, but rather only that the observer be in a position to check for the presence (absence) of the effect following the presence (absence) of the presumed cause in (ideally) the widest possible range of surrounding circumstances. This suggestion would have us view experimental interference as an expedient, by which we put ourselves in a position to observe the behavior of the factors in a range of circumstances which wouldn't ordinarily come into our ken... (McCann 1978, p. 90-91.)

The account of how experiments can reveal causal relations, that is based in **ICI**, understood as a theory of intervention, differs radically from the one described by McCann. Because it implies that, if we passively observe a naturally occurring correlation, and then manipulate one of those correlated factors and observe *precisely the same things as before*, with respect to all external events—that is, there is "invariance under intervention"—then we are nevertheless in a radically different epistemic situation after having performed the manipulation, that enables us to gain new causal knowledge about the system. And this is explained, not by the free actions of agents being ultimately independent of all natural causes, but by the physical and causal properties that characterize agents. Agents are, in virtue of having these natural properties, interventions waiting to happen.

CHAPTER 9

Conclusion

9.1. Assessing manipulationism

This work began with a question about what role manipulations have in our relationship to causation. It seemed natural to orient such an investigation around the explicitly manipulationist takes on causation, displayed in the theories we have reviewed here, since these theories of causation, in contrast with historically more mainstream ones, give manipulations an essential role in our understanding of causation and causal concepts. Moreover, the investigation would eventually focus on James Woodward's interventionist account, which has attracted mainstream interest in a way the earlier manipulationist theories have not.

In the broadest terms, the *manipulationist intuition* seems to me to be that agency and manipulation play an indispensable part in our relationship to causation, including the formation of causal concepts. I take it that this can hold true for both causal realists and anti-realists. If we are causal anti-realists—in the sense that what is a cause of what is not fully determined by objective, natural facts, that are independent of the constitutions and beliefs of agents—then the intuition can provide substance for an analysis of the cause concept, that preserves its meaningfulness, as it were, in the face of the anti-realist arguments. The belief that causation is necessarily related to free agency has been around at least since Thomas Reid's *Essays on the Active Powers of Man* (Reid 2010, Essay I, ch. V), where it is made explicit, and it may have been implicit in causal concepts long before that (Collingwood 1940, ch. XXXII). The details of such an account may differ depending on the author's views on such things as "projection" (Humean or Kantian?), as we have seen in the reviewed theories. Some of them are moreover difficult to classify—e.g. Gasking's proposal, due to the sparsity of the source material, and von Wright's because of a tension between his manipulationism and his compatibilism.

Regardless of these details, I take it that the proposal will not go through unless the analysis delivers an interpretation that conforms to our most important causal judgments. The argument in section 8.2.2, to the effect that manipulationism doesn't provide a sufficient condition for causation—something already noted by Woodward—addresses this concern. If a theory implies that a manipulation can't introduce confounding correlations in the outcome, then relative to important insights about causation and manipulations, mainly acquired in the context of scientific experimentation, that is a *reductio*. We know that they can. What is missing in the manipulationist account, then, is the

condition that the manipulation must be "surgical." That resolving this issue is not a matter of simply adding the surgicalness condition to the theories, I take to be shown in the subsequent sections of that chapter.

Two things (at least) change with James Woodward's proposal in *Making Things Happen*. First, the conditions for the presence of a causal relation in the context of a manipulation are amended so as to be, in a sense, sufficient (at least relative the understanding provided by scientific causal inference theories). Second, Woodward defends causal realism. Both of these differences have important implications.

First, I've named several different ways in which a manipulationist theory can be circular in how it purports to define causation. Some of them can, in theory, be avoided (suspending judgment on whether the resulting theory is then adequate in other respects). But in a sufficient set of manipulationist conditions on causation, some circularity appears unavoidable, in particular when we must require that the manipulation doesn't cause the outcome unless by way of its direct result (surgicalness), and that it doesn't share a cause with the outcome (exogeneity).

Due to these circularities, Woodward's theory can't explicitly define causation, and I think they are too tight for something like a "connective analysis" in Strawson's sense to be plausible. This result is by extension applicable also to the older manipulationist theories, in that, if their conditions for causation are made sufficient, then the same circularity, with the same consequence for the interpretation of causal claims, can be expected to occur. Woodward's conditions also prove to be too weak to determine a unique causal relation, when the theory is treated as an inductive or an implicit definition of direct causation.

I'm aware that Woodward may not, especially more recently, mean that interventionism provides for the meaning of causal claims in any of the ways I have assumed here. But every indication in *Making Things Happen* points in the direction of an *explication*, or possibly a *connective analysis*, and I'm not aware of any other account of what Woodward has in mind, as regards this semantic goal specifically. (That the goal is to provide a *functional* concept of causation isn't really an alternative—that's been an aspect of explication from its Carnapian beginnings.) My aim, then, may conceivably be wide of the mark in this evaluation of Woodward's theory. But if it promotes a more informative account of how, precisely, interventionism is meant to explain the meaning of causal claims, or be a theory of causation generally, then I'll consider the effort to have been justified.

Second, the fact that the interventionist theory can't define a unique causal relation proves to make causal realism—understood as implying that the causal relation has a certain determinate extension in the real world—a commitment above and beyond the commitment to the interventionist theory. To return to the semantic project, the question is then: if the interventionist conditions don't deliver a determinate, real causal relation, and we commit to one all the same, why should we *not* employ that relation directly in our interpretation of causal claims?

9.2. What is illuminated?

I don't think, then, that interventionism can be a theory of causation. However, I agree with the manipulationist intuition, that I described above, as well as with Woodward's causal realism. I also think that interventionism's roots in theories for causal inference are sound. In the introduction to this monograph I characterized manipulationist theories in terms of a certain implication, namely that there is a possible manipulation or intervention on every cause. I have called this condition **NEC**. I have argued that interventionism's sufficient condition, expressed roughly by **IC 1**, must be taken to state something true about causation as it is in the world (at least when the Causal Markov Condition is satisfied), for the theory to underwrite sound causal inferences. Its necessary condition **NEC**, on the other hand, cannot be interpreted in this way. If **NEC** says something true, then it is rather about certain *formal operations* on variables, or *precisifications* of the meaning of certain variables, and not about certain kinds of *causal factors* in causal systems. Given the commitment to a real, determinate causal relation that causal realism entails, I think that **IC 1** can be retained, together with a primitive causal relation C. I call the resulting theory **ICI** and suggest that it's not a theory of causation—and it is in particular, absent **NEC**, not a manipulationist or interventionist theory of causation—but rather a theory of *intervention*. What remains of the theory in **ICI** are then, I think, precisely its roots in causal inference theory, under a certain understanding of what such theories are about.

I have argued that **ICI** can contribute substantially to our understanding of *manipulation*, and specifically to its epistemic role in causal knowledge acquisition. It can, in this way, substantiate the intuition that I think drives the manipulationist proposals, i.e., the idea that manipulations are essentially involved in our *acquaintance* with causation, and therefore also in the historical *formation* of our causal concepts (which is not to be confused with their *content*).

Employing interventionism as a theory of intervention, rather than a theory of causation, and applying it to human manipulations, suggests an epistemic role for manipulation, and thus for experimentation, that is radically different from the one associated with historically more mainstream views, that mainly came out of regularism and its singular focus on passive observation. In the last part of the dissertation, I have sketched an account of non-inferential causal knowledge acquisition, that relies on this theory of intervention, and according to which our epistemic situation changes depending on whether we are personally performing a manipulation or not, all else being equal. I suggest that an account such as this could explain how we have access to some simple causal facts, that can then be employed in overt causal inferences. From one perspective, what I describe is very similar to Menzies's and Price's claims about directly experiencing instances of "bringing about"—the difference being just that, on my view, this is a purely epistemic affair, with no implications for what causation is, or what causal claims mean. (I.e., what C_R is *really* all about.) This account, in addition, is not just compatible with, but depends on, a naturalistic view of agents as causal systems in themselves, with certain

essential causal and physical properties. It has nothing, then, of the old treatments of free agency as something apart from the natural world. Finally, if it is correct that we can directly experience the phenomenon of causation when we perform manipulations, then taking the causal relation as a primitive can also seem more acceptable from an empiricist perspective: "cause" is then an *empirical notion*, that doesn't require a definition or analysis to be meaningful.

To summarize, manipulationism and interventionism fail as theories of causation in the sense of theories of the meaning of causal claims, under the assumption that standard definition theory gives us the conditions under which such an enterprise succeeds. What I propose as the alternative, is to eliminate the manipulationist condition **NEC** from the theory, leaving **IC 1**, and adding a theoretically primitive causal relation C. The resulting theory **ICI** is an illuminating theory—but the object illuminated is primarily *intervention*, and by extension manipulation in the context of successful causal experiments. To this we must add a naturalistic account of how we can come to have some knowledge of the causal relation. I think **ICI** can be useful in this respect, and that this in fact reveals the essential, epistemic role that manipulation has, in our acquaintance with causation.

I want to add in a final note, that Woodward's idea of *invariance*, that plays a crucial part in his theory of causal explanation, and that I have not discussed at all, fits extraordinarily well with an issue I brought up in the introduction to this monograph. There I mentioned that the question that originally led me to this investigation was how regularities that we are inclined to explain in terms of (for example causal) laws differ from other regularities that we may be observing. The context was Armstrong's defense of modally strong laws as explainers of regularities. It seems to me that "invariance under intervention" is just the property of a regularity that I was looking for.

9.3. Questions for future research

First, the traditional questions about the causal relation remain, on my view. What kinds of things (and how many things) does it relate? Is it determined by non-causal facts? How does it connect to fundamental physics? Etc. They are to a large extent independent of the interventionist theory for causal inference, but not completely.

Second, we may take our acquired understanding of an intervention as a starting point for an investigation into the precise epistemic role of manipulations and experiments. My sketch of such an account at the end of the last chapter is a sketch, and what is sketched is in large parts a scientific theory, not a philosophical one. A proper account depends in particular on a proper understanding of the sense of agency, its prevalence, properties, and triggers, and thereby on the related research in empirical psychology. (But it also depends on a proper account of sense-based *knowledge*—a question for philosophy.) One aspect for which I have found no experiments, and that relates importantly to considerations of the sense of agency as a source of *new* causal knowledge, is how the sense operates in relation to effects of manipulations that are wholly unexpected by the agent. Many other specific questions, of both a scientific

and a philosophical nature, would doubtless appear if this sort of account is pursued. There should also be close connections to existing empirical research into causal learning in general.

Within philosophy, this account would sort among our theories about how we may directly experience causation, that have more commonly focused on passive sense experiences, and observation in particular. It may be that the experience of causation under manipulation is more epistemically basic to agents than the identification of causes in passive observation, which may in that case validate the manipulationist intuition to a considerable degree (but only under an epistemic interpretation of that intuition).

Precisely how causal knowledge acquired experientially under manipulation fits into the general scheme of causal inference need also be made more precise. What type of causal facts are thus acquired, and what does the next step need to look like, as we enrich our causal understanding of some nearby part of the world?

Third, it seems to me that our identification of the proper sense of "free" in "free action," in the context of manipulationist theories, may have intriguing implications for a compatibilist understanding of free will (or at least free agency). We noted that the sense in which a manipulation must be free, for the manipulation to do its epistemic job of revealing causal relations, was not "free" as in "voluntary," but "free" as in "causally exogenous." It is, in a sense, the fact that the manipulation is exogenous relative to the manipulated system, and to the outcome in particular, that makes the agent an *agent* relative to this local context. I moreover argued that we can believe that manipulations by agents regularly *are* exogenous in the required way, on perfectly naturalistic and scientifically non-mysterious grounds, that relate to the complexity of the agent's decision mechanism, its causal isolation from the external physical environment, and the dilution of causal influences along a path over time, due to causal noise. Spinoza claimed that mankind has an illusion of free will, because she knows her actions but not their causes (1677). The latter is indeed implied in this account. But it is moreover the very same situation that implies this, that also *makes* the agent an agent, relative to the local manipulated system. Shades of a compatibilist theory of (local) free agency? More to come, perhaps.

Svensk sammanfattning

Manipulationsteorier och interventionsteorier om kausalitet. Hur relationen mellan orsak och verkar ska förstås är ett filosofiskt problem som går tillbaka till antiken. Fokus i filosofiska teorier om kausalitet har varierat kraftigt, från till exempel antagandet att en *förklaring* till en händelse omnämner dess orsak, till *fysiska interaktioner* mellan orsaksrelaterade händelser, till en essentiell eller begreppslig relation mellan orsakande och *agens*. I modern tid har diskussionen vanligen präglats av en empiristisk filosofi, där orsakssamband grundats i universella regulariteter eller korrelationer. Regulariteter och korrelationer har betraktats som empiriskt legitima utgångspunkter för en analys, i kraft av att de kan observeras, i motsats till en mer metafysiskt substantiell orsaksrelation, som till exempel medför någon sorts objektiv *nödvändighet* i orsakandet. Inom detta empiristiska ramverk har vissa problem stått i förgrunden, som att etablera orsaksrelationens asymmetri, skilja korrelationer som förklaras av ett orsakssamband från de som istället förklaras av en gemensam orsak, eller att ge en förklaring till den nödvändighet med vilken en orsak åtminstone *tycks* frambringa sin verkan. David Humes empiristiska argument har vanligen varit utgångspunkten för detta synsätt (1888; 1902). Ett alternativ till Humes analys av kausalitet har istället knutit orsakssamband, eller våra kausala begrepp, specifikt till *agens* och *manipulationer*, eller våra trosföreställningar om sådant. Tanken går tillbaka åtminstone till Thomas Reid (2010), och har försvarats av en minoritet av filosofer sedan 1930-talet. Denna syn på kausalitet har blivit mer populär bland filosofer och andra forskare under de senaste årtiondena, främst av två orsaker. Den ena är utvecklingen av nya matematiska verktyg inom vetenskapen för att identifiera och kvantifiera orsakssamband, gjorda av bland andra Peter Spirtes, Clark Glymour och Richard Scheines, samt Judea Pearl, där man i dessa metoder förlitar sig på *interventioner* i en teknisk bemärkelse för att definiera en kausal effekt (Spirtes et al. 2000; Pearl 2009). Den andra orsaken till manipulationismens ökande popularitet har varit James Woodwards teori om kausalitet, som till stor del inspirerats av dessa vetenskapliga matematiska verktyg (2003).

Den här monografin undersöker manipulationers roll i teorier om kausalitet. Frågan kan i första ledet delas upp i en historisk—eller deskriptiv—del, samt en normativ del. Vi kan alltså fråga oss hur manipulationer faktiskt använts för att förklara eller belysa någon aspekt av orsakande, och precis vilka aspekter dessa förklaringar fokuserat på, och vi kan också försöka förstå närmare vilken roll manipulationer generellt *kan* ha i detta avseende. Vi kan göra en ytterligare uppdelning mellan å ena sidan manipulationistiska förklaringar

av kausalitetens *metafysik* eller innehållet i kausala *begrepp*, och å andra sidan manipulationistiska förklaringar av hur vi förvärvar *kunskaper* om orsaker, det vill säga *epistemiska* förklaringar. All de teorier som behandlas i monografin berör kausalitetens metafysik eller semantik—det är detta som normalt antas göra dem just till manipulationistiska teorier om kausalitet. Men frågan om manipulationers epistemiska roll är av stor vikt i min egen diskussion om hur vi som individer formar en bekantskap med kausalitet betraktat som ett naturligt fenomen, vilket i sin tur relaterar på ett viktigt sätt till de intuitioner som varit drivande hos manipulationister.

En manipulation är här, i stora drag, en typ av handling, utförd av en agent, och som i någon bemärkelse är fri(villig). En manipulation har alltid vad vi kan kalla en "direkt konsekvens" ("direct result"). För att låna ett exempel från von Wright, om vi öppnar ett fönster för att vädra rummet, så benämner vi händelsen att fönstret öppnas som manipulationens direkta konsekvens. Händelsen att rummet vädras kallar vi istället ett "utfall" ("outcome"). Det är orsaksrelationen mellan dessa två händelser, det vill säga manipulations direkta konsekvens och utfallet, som manipulationsteoriernas villkor för kausalitet tillämpas på.

Det är lätt att tillerkänna manipulationer en särskild roll i förklaringar av hur vi förvärvar kausal information—i synnerhet i samband med experiment—utan att därmed förespråka en manipulationsteori om vad kausalitet är, eller om orsakspåståendens semantiska innehåll. Inom ramarna för avhandlingen definierar vi därför en manipulationistisk teori om kausalitet i termer av en specifik implikation från teorin: teorin implicerar att, om något A är en orsak, så kan antingen A manipuleras (vara en direkt konsekvens av en manipulation), eller så har A en viss sorts relation till något manipulerbart. Manipulerbarhet, eller rätt sorts relation till något manipulerbart, är alltså ett nödvändigt villkor för att något ska vara en orsak, i en teori som vi här kallar manipulationistisk. "Manipulerbar" kan förstås på olika sätt i olika teorier, och till exempel syfta på manipulerbarhet *i praktiken*, eller på något svagare såsom "manipulerbar *i princip*". "Manipulation" ersätts därtill med "intervention" i den nyaste av de teorier vi behandlar.

Avhandlingen är orienterad runt en kronologisk genomgång av fem olika manipulationistiska teorier, som presenterats av R G Collingwood (1938; 1940), Douglas Gasking (1955), G H von Wright (1971; 1973; 1974), Peter Menzies och Huw Price (1993) samt James Woodward (2003). Av dessa är det Woodwards teori som fått störst genomslag bland filosofer och andra forskare. Denna teori är starkt influerad av vissa matematiska metoder för att uttrycka kausala hypoteser och härleda konsekvenser från dessa, såsom *Structural Equation Modeling* (SEM), *Structural Causal Modeling* (SCM) och även i vissa avseenden *Potential Outcomes Framework* (ex Spirtes et al. 2000; Pearl 2009; Holland 1986). Nyckelbegreppet i dessa sammanhang är inte "manipulation" utan "*intervention*". Inom ramarna för SEM, SCM och Woodwards teori definieras en intervention på en kausal faktor X i termer av sina orsaksrelationer till X och till andra delar av Xs system, såsom ett visst utfall Y. En manipulation kan därmed misslyckas med att uppfylla villkoren för att vara en intervention, och en intervention kan i sin tur vara något annat än en handling utförd av en agent. Begreppen är

alltså överlappande, men inte synonyma, och inget av dem innefattar heller till fullo det andra. Men för att interventioner är en sorts generalisering av manipulationer (sprungen i huvudsak ur tanken på ett idealiserat experiment), och för att Woodward själv kallar sin teori för en manipulationsteori i *Making Things Happen*, har jag använt "manipulationsteori" som den övergripande etiketten. Detta trots att problem och frågeställningar i flera viktiga avseenden skiljer sig mellan de tidigare manipulationsteorierna och de som är formulerade i termer av interventioner.

I övergripande drag kan en manipulationsteori uttryckas enligt följande:

M: *A* är en orsak till *B om och endast om (i) A* kan vara föremål för en manipulation/intervention M och A och B samvarierar när M sker, eller *(ii)* A och B har en särskild relation R till några A' och B' som uppfyller villkor *(i)*.

Vi kan läsa ut "$R(A, A')$" till exempel som "A är en händelse av samma typ som A'", där specifik typindelning kan ske på olika sätt i olika teorier.

Precis vad A och B—det vill säga orsaksrelationens relata—är skiftar mellan olika teorier, men kan indelas grovt i händelsetyper och faktorer. En faktor är en egenskap hos ett kausalt system som kan anta ett av ett flertal olika värden, och symboliseras av en variabel i en modell. Ett speciellt men vanligt fall är när en binär variabel "X" står för en händelse, och "$X = 1$" indikerar att händelsen inträffar och "$X = 0$" att den inte gör det. Det är hos de senare interventionsteorierna som faktorer kan utgöra kausala relata, och dessa kallas vanligen rätt och slätt "variabler". De äldre teorierna relaterar snarare händelsetyper, där "typ av händelse" kan förstås på olika sätt beroende av teori (till exempel explicit som en typ av förändring i ett systems tillstånd mellan två tidpunkter). En händelsetyp kan också definieras helt generellt med utgångspunkt från en faktor, och då uttryckas som en tilldelning av ett värde till en variabel: "$X = k$"—så att orsakande mellan händelser ges en tolkning även i detta sammanhang.

Om manipulationen eller interventionen M per definition ges, eller generellt kan antas ha, vissa specifika egenskaper så kan under vissa antaganden det *tillräckliga villkoret* i **M** styrkas på logiska grunder. Om vi ignorerar villor *(ii)* i **M** för att förenkla resonemanget, så är detta tillräckliga villkor för kausalitet, som följer ur **M**:

$\mathbf{T^M}$: Om A kan vara föremål för en manipulation/intervention M och A och B samvarierar när M sker, så är A en orsak till B.

Vi kan nu göra följande antaganden. Först, Reichenbach's *Common Cause Principle* (**CCP**): om A och B samvarierar så är antingen den ena en orsak till den andra, eller så har de en gemensam orsak. Detta är bara rimligt om samvariationen mellan A och B när M sker inte är en artefakt av att vi använt oss av ett litet stickprov. Vi avser alltså en samvariation i "populationen". Sedan antar vi att händelsen M uppfyller villkoren **MV** nedan.

MV

(1) A har ingen annan orsak än M.
(2) M är inte en orsak till B annat än möjligen *via* A.

(3) M delar inte någon orsak med B vars påverkan på B inte förmedlas av A.

2 benämner vi som villkoret att M utgör ett "*kirurgiskt*" ingrepp på A ("the surgicalness condition"), och 3 innebär att M är kausalt *exogen* i förhållande till B ("the exogeneity condition").

Under dessa antaganden följer det tillräckliga villkoret $\mathbf{T^M}$ för kausalitet. Att A är en orsak till B följer under dessa antaganden från förledet i $\mathbf{T^M}$, eftersom samvariationen mellan A och B, per **CCP**, då måste bero på att A är en orsak till B eller tvärtom, eller förklaras av att de har en gemensam orsak, och om M dessutom uppfyller villkoren **MV** så kan inte B vara en orsak till A, och deras samvariation kan inte heller förklaras av en gemensam orsak. Det återstår då bara att A är en orsak till B. Att $\mathbf{T^M}$ följer från **CCP** tillsammans med dessa antaganden om manipulationer kan ses som tillräckligt för ett försvar av $\mathbf{T^M}$ betraktad som en *inferensregel*, som tar oss från vissa fakta om orsakanden och samvariationer till ny någon kausal information. Detta är just situationen i de teorier om kausala inferenser som nämns ovan. Där är $\mathbf{T^M}$ en sådan inferensregel, som kan visas vara giltig under antagandet att en viss kausal modell satisfierar *Causal Markov Condition*. Causal Markov Condition är, i stora drag, **CCP** betraktad som en kontingent egenskap hos vissa kausala modeller. Men **CCP** implicerar inte **M**, och inte heller tvärtom, och rättfärdigandet av $\mathbf{T^M}$ sedd som en kausal inferensregel är inte tillräckligt för att visa att vi därmed har en teori om vad kausalitet är eller vad orsakspåståenden betyder, eftersom inferensen kräver att vi har viss kausal information i premisserna.

Ett fokus i avhandlingen ligger på hur de olika teorierna försvarar detta tillräckliga villkor $\mathbf{T^M}$ för kausalitet, bland annat genom att explicit eller implicit tillskriva manipulationen eller interventionen M egenskaperna som anges i **MV**. Dessa egenskaper anges explicit i Woodwards *definition* av en intervention, och de återfinns som sagt, på ett mer informellt sätt, i de matematiska ramverken för kausala inferenser som till stor del inspirerat Woodward. Att "fria" (i betydelsen "frivilliga"/"voluntary") manipulationer har *vissa* av dessa egenskaper är snarare ett mer eller mindre implicit antagande i flera av de äldre manipulationsteorierna. Detta är en avgörande skillnad mellan interventionismen och den tidigare manipulationismen.

Ett annat av avhandlingens fokus ligger på en teoris rättfärdigande av det *nödvändiga villkoret* för kausalitet som impliceras av **M**, och som här identifierar en teori som manipulationistisk. Detta säger alltså:

> **$\mathbf{N^M}$:** Om A är en orsak till B så kan A vara föremål för en manipulation/intervention M, och A och B samvarierar när M sker.

Även i detta avseende skiljer sig de äldre teorierna från Woodwards, främst i det att de äldre teorierna till någon grad är antirealistiska med avseende på kausalitet, och därför förespråkar ett uttryckligt *antropocentriskt* orsaksbegrepp, som är nära kopplat till agenters (kanske föreställningar om sin) förmåga till, och intressen av, kontroll. ($\mathbf{N^M}$ kan då tolkas som en utsaga till exempel om agenters trosföreställningar.) Till skillnad från detta är Woodward kausal realist och vill därför undvika antropocentriska implikationer.

I ljuset av dessa frågor har det varit nödvändigt att försöka förstå närmare vad förespråkarna för dessa teorier avsett att uppnå med teorin. Vad alla har gemensamt är att de avser att ange innehållet i kausala begrepp och orsakspåståenden. Det rör sig alltså om en *begreppsanalys* av något slag. Även om det varit vanligt av avsvära sig explicit metafysiska ambitioner i dessa sammanhang är begreppsanalysen mycket närliggande en teori om vad kausalitet *är* (och distinkt från en epistemisk teori om hur vi förvärvar kausala *kunskaper*). Vi kan identifiera fyra mer specifika mål som dessa teorier kan ha, i olika kombinationer och till olika grad:

(1) Att förklara intrycket av *nödvändighet i orsaksrelationen*, i ljuset av traditionella empiristiska invändningar mot att någon sådan nödvändighet existerar i den objektiva, fysiska världen.
(2) Att försvara *orsaksbegreppens meningsfullhet*, återigen i ljuset av traditionella empiristiska invändningar.
(3) Att begreppsligt etablera *orsaksrelationens asymmetri*, särskilt i ljuset av naturlagarnas symmetri.
(4) Att begreppsligt skilja orsakssamband från fall där en korrelation förklaras av en *gemensam orsak* ("spurious correlation").

Av de teorier som granskas är det bara Collingwoods och von Wrights som explicit tar sig an 1, medan alla på ett eller annat sätt omfamnar de övriga målen, utom Collingwood som inte säger något om 3 eller 4.

Tidigare invändningar mot manipulationism. Peter Menzies och Huw Price identifierar och bemöter fyra invändningar mot manipulationsteorier om kausalitet, som de menar har varit vanligt förekommande i litteraturen (Menzies and Price 1993):

(1) Många verkliga orsaker *kan inte i praktiken manipuleras*, och manipulationsteorierna kan därför inte inkludera dessa.
(2) Manipulationsteorierna är alltför *antropocentriska* (eller *-morfa*).
(3) I sitt fokus på manipulationers roll *förväxlar manipulationsteorierna ontologi och epistemologi.*
(4) Då "manipulation" vanligen betraktas som ett kausalt begrepp i sig är teorierna *cirkulära* i sin definition av kausalitet.

Invändning 1 har, så vitt jag kunnat se, inte varit vanlig i kritiken mot manipulationsteorier. Tvärtom tycks alla förespråkare göra reda för detta förhållande på något sätt. Collingwood förnekar helt enkelt att något som inte kan manipuleras (i praktiken) med rätta skulle kunna kallas en orsak. Gasking och von Wright—och Menzies och Prices själva—utvidgar orsaksrelationen till ej manipulerbara händelser genom någon relevant relation till manipulerbara fenomen, på det sätt som angavs i **M**. (Orsaken är alltså i någon bemärkelse av samma typ som något manipulerbart.) Woodward (som skriver senare än Menzies och Price) hävdar snarare att alla orsaker *är* föremål för en möjlig intervention, där "möjlig" förstås som något svagare än "praktiskt möjlig" och till och med "fysiskt möjlig". (Det härrör snarast från idén "möjlig i princip" som är vanligare i den vetenskapliga än i den filosofiska litteraturen.)

Invändningar 2–4 har förekommit upprepade gånger i kritik mot manipulationismen (ex Rosenberg 1973; Mackie 1980; Hausman 1986). Jag argumenterar i avhandlingen för att träffsäkerheten hos dessa invändningar ändå kan ifrågasättas, till olika grad.

Att de tidigare teorierna är *antropocentriska* (invändning 2) tycks vara ett implicit *mål*. Avsikten med dessa tidigare teorier är delvis att försvara ett meningsfullt orsaksbegrepp givet att kausalitet *inte* är en del av den agentoberoende fysiska världen. Manipulationisternas strategi kan, i grova drag, jämföras med Humes, när Hume förklarar intrycket av *kausal nödvändighet* i termer av en förväntan att en händelse av typ B ska följa en händelse av typ A, som uppstår hos en individ om B-händelser alltid följt A-händelser i hennes erfarenhet. De tidiga manipulationisterna skiljer sig från Hume, såtillvida att förklaringen av det psykologiska fenomenet är annorlunda och mer komplicerad, men precis som hos Hume är den kausala nödvändigheten för dem ett mentalt fenomen. Antropocentrism kan därför åtminstone inte vara ett *internt* problem i dessa teorier—utan en konflikt måste då snarast lokaliseras till frågan om kausal realism i sig. Om kausal realism inte är ett antagande verkar också invändning 3 bli mer komplicerad.

Invändning 4 utgår från antagandet att dessa teorier avser att *definiera* orsaksrelationen, i någon bemärkelse. Detta verkar vara ett rimligt antagande om det är fråga om någon form av begreppsanalys. I min egen diskussion har jag fokuserat på dessa teorier just i deras egenskap av teorier om meningen hos kausala begrepp. Cirkularitetsinvändningen har därför en särskild betydelse i avhandlingen. I den idag mest populära manipulationsteorin, James Woodwards, framstår också cirkulariteterna i ett särskilt ljus, då Woodward på sätt och vis omfamnar dem, åtminstone såtillvida att han benämner sin teori som "icke-reduktionistisk". I avhandlingen skiljer jag mellan tre typer av cirkularitet som kan förekomma i de manipulationistiska teorierna (kap 8).

Cirkularitet$_1$*:* en manipulation av A sägs eller antas i sig vara en *orsak* till A.
Cirkularitet$_2$: teorin anger som villkor för att A ska vara en orsak till B att vi *frambringar* (el dyl) B genom att manipulera A, där "frambringa" är en kausal term.
Cirkularitet$_3$: teorin inkluderar något av de kausala villkoren i **MV**, ovan.

(Alla dessa typer av cirkularitet har berörts mer eller mindre explicit i tidigare litteratur.) Jag visar att ingen av Collingwoods, Gaskings eller von Wrights teorier lider av cirkularitet$_1$. Det är en konsekvens av de villkor de uppställer att en manipulation *inte* kan sägas orsaka sin direkta konsekvens. Collingwood beskriver en orsak som ett *medel* för en agent att uppnå en viss verkan, och anger uttryckligen att en manipulation *inte* är ett medel för att uppnå manipulationens direkta konsekvens, och därmed är manipulationen inte heller den direkta konsekvensens orsak. Ett liknande resultat är en, möjligen oavsiktlig, logisk konsekvens av Gaskings villkor på orsaksrelationen. Von Wright är mest explicit i sitt argument mot att denna typ av cirkularitet föreligger i hans teori, men hans särskilda sorts *kompatibilism* leder ändå till en tvetydighet i detta sammanhang. Menzies och Price själva avser att undvika denna typ av cirkularitet genom att grunda kausala begrepp i begreppet "bring about", som

de menar inte är kausalt, och i någon bemärkelse föregår kausala begrepp hos agenter.

Huruvida cirkularitet$_2$ förekommer i en teori kan bero på hur författaren väljer att formulera teorin, från ett tillfälle till ett annat. Istället för att säga att vi *frambringar* B genom att manipulera A, kan vi säga att B *sker* (eller förändras, om B är en faktor) när A manipuleras. Vi hänvisar då till en samvariation istället för direkt till ett kausalt frambringande. Cirkularitet$_2$ tycks därmed lätt att undvika i en manipulationsteori. Under den senare formuleringen öppnas dock för möjligheten att samvariationen mellan A och B under manipulationen eller interventionen M har en annan förklaring än att A är en orsak till B. M kan till exempel i sig vara en sådan händelse att den har en gemensam orsak med B, eller orsakar B oberoende av A. (Till exempel om vi skruvar på A med handen, men samtidigt råkar påverka B med armbågen.) Det är sådana möjligheter som utesluts av villkor 2 och 3 i **MV**. När dessa villkor inkluderas i teorin introduceras istället cirkularitet$_3$.

Min analys rör i första ledet vad dessa förhållanden har för konsekvenser för en manipulationistisk teori som avses förklara meningen hos kausala begrepp och orsakspåståenden, genom att ge en *definition* av orsaksrelationen. Jag visar att ingen av dessa teorier kan lyckas på denna punkt, åtminstone inte under standardantaganden om vad som krävs för att en definition ska vara framgångsrik (avsnitt 8.2–8.4). Men i nästa led argumenterar jag också för att en delmängd av Woodwards interventionistiska teori, tillsammans med vissa ytterligare antaganden, kan bidra till vår förståelse av *manipulationer*, och av manipulationers roll i vår personliga relation till kausalitet (avsnitt 8.7). I detta argument finns två slutsatser som tycks bekräfta manipulationismens drivande intuitioner, men utan att därmed ge manipulationer någon roll i kausalitetens metafysik eller orsaksbegreppens semantik. För det första kan vi förvärva kausal information genom *direkt erfarenhet*, just när vi utför manipulationer. Jag kopplar en förklaring av detta förhållande till en särskild typ av upplevelse, vår "*sense of agency*". Detta är upplevelsen av att ha orsakat en händelse genom en handling—något det på senare tid utförts empirisk psykologisk forskning om. Jag knyter detta komplexa sinne till villkoren i **MV**, tillämpade på faktiska manipulationer utförda av människor. Processen genom vilken vi förvärvar kausal information via direkt erfarenhet när vi utför manipulationer antas vara felbar, men nog tillförlitlig för att vara epistemiskt viktig. För det andra är, i kraft av detta förhållande, "orsak" ett *empiriskt begrepp*, som därför, ur ett empiristiskt perspektiv, inte *behöver* analyseras eller definieras i icke-kausala termer för att vara meningsfullt.

Manipulationismen misslyckas som en teori om meningen hos orsakspåståenden. Här beskriver jag avhandlingens argument och slutsatser, och mina antaganden, i korthet. Jag antar som sagt att manipulationsteorierna avses ge oss meningsinnehållet i orsakspåståenden. Jag antar därtill att standard definitionsteori förser oss med villkoren under vilka detta är fallet (Belnap 1993; Lewis 1970).

Manipulationsteorins tillräckliga villkor för att A ska vara en orsak till B är som vi såg:

> **$\mathbf{T^M}$:** Om A kan vara föremål för en manipulation/intervention M och A och B samvarierar när M sker, så är A en orsak till B.

För att $\mathbf{T^M}$ ska vara sant måste villkoren i **MV** vara uppfyllda av M. De äldre manipulationsteorierna (alla utom Woodwards) inkluderar inte villkor 2 i **MV**. Konsekvensen av detta är att teorin identifierar A som orsak till B även då den rätta förklaringen till deras samvariation är att de har M som en gemensam orsak. Teorins villkor är alltså inte tillräckligt för att vara deskriptivt adekvat. Om villkor 2 i **MV** läggs till en sådan teori blir den istället cirkulär, eftersom 2 innehåller predikatet "_ är en orsak till _". Kravet från definitionsteori, att definiendum kan elimineras överallt till förmån för dess definiens i teorin, är då inte tillfredsställt, och teorin misslyckas därför även nu med att ge meningsinnehållet i orsakspåståenden. Detta argument beaktar bara hur en definition av orsakande explicit har formulerats. Analysen av huruvida en teori av denna typ kan ge oss en definition av orsaksrelationen behöver därför fördjupas, och detta sker i samband med Woodwards teori.

I Woodwards interventionistiska teori ersätts "manipulation" med "intervention". En intervention I^{XY} på en faktor X med avseende på en annan faktor Y är en händelse som per stipulation uppfyller villkoren i **MV** med avseende på X och Y. Att denna teori inte utgör en lyckad explicit definition av "_ är en orsak till _", på grund av att **MV** introducerar cirkularitet har vi redan sett. Jag visar därefter att teorin inte heller är en framgångsrik induktiv eller implicit definition av orsaksrelationen. Det sistnämnda innebär att vi modellteoretiskt kan konstatera att teorins villkor inte är tillräckliga för att bestämma en unik extension för orsaksrelationen, givet en specifik tolkning av teorins icke-kausala grundspråk. Jag förlitar mig på några informella tillämpningar av Padoas metod för att visa detta (Craig 1956). Inte heller kan teorin sägas ge en plausibel *approximation* av meningen hos orsakspåståenden, då teorin alltid är kompatibel med att X inte är en orsak till Y, för vilka X och Y som helst, och även med att inget alls är en orsak till någonting, detta oavsett vilka ickekausala förhållanden (till exempel korrelationer) som föreligger.

Den sammantagna slutsatsen är alltså att, givet att en teori framgångsrikt ger meningsinnehållet för ett begrepp endast om den definierar detta begrepp i enlighet med standard definitionsteori, så kan en manipulationsteori av de slag vi här har haft att göra med inte ge oss meningen hos orsakspåståenden.

Jag diskuterar därefter ett sätt på vilket villkoren i Woodwards interventionsteori kan stärkas. Jag visar att även om den resulterande teorin framgångsrikt bestämmer en unik extension hos orsaksrelationen, givet alla för teorin relevanta ickekausala fakta, så kommer denna manöver i konflikt med Woodwards kausala realism. Den kan inte heller anses få stöd från den vetenskapliga litteraturen om kausala inferenser, som Woodward till viss del förlitar sig på.

Jag avslutar diskussionen om manipulationsteorierna med att föreslå att villkoren $\mathbf{T^M}$ och **MV**, som de förekommer i den interventionistiska teorin, trots allt utgör en plausibel *restriktion* på orsaksrelationen. Vi kan därmed anta som primitiv en reell orsaksrelation C_R och betrakta $\mathbf{T^M}$ och **MV** tillsammans som en restriktion på C_R. Interventionsteorins nödvändiga villkor

$\mathbf{N^M}$ för kausalitet förkastas, för att det inte kan ges en rimlig realistisk tolkning, och jag kallar C_R+$\mathbf{T^M}$+**MV** för **ICI**. **ICI** är inte en manipulationistisk teori om kausalitet, eftersom $\mathbf{N^M}$ inte är en implikation. Det är inte en teori om vad kausalitet är, eller om meningen hos orsakspåståenden, överhuvudtaget eftersom den saknar ett nödvändigt villkor för kausalitet och tar en orsaksrelation som teoretiskt primitiv. Men det är *en teori om intervention*. Den anger villkoren under vilka en manipulation kan ge oss ny kausal information och förklarar varför just dessa är villkoren. Jag menar att detta är den *reella* delen av Woodwards teori, och att den kan spela en central roll i vår förståelse av kausala experiment, av vår personliga bekantskap med kausalitet som fenomen, och därmed också av de intuitioner som tycks driva manipulationistiska försök att förklara vad kausalitet är eller vad orsakspåståenden betyder. Detta utan att vara en teori om kausalitet eller om orsakspåståendens mening.

Vår bekantskap med kausalitet. ICI förser oss med villkoren under vilka ett orsakssamband mellan X och Y följer från en samvariation mellan X och Y under en manipulation M av X. (Nämligen när manipulationen är en intervention.) I den avslutande delen av monografin försvarar jag tesen att när en manipulation M av X är en frivillig handling och X samvarierar starkt med en annan faktor Y, då är det *sannolikt* att M uppfyller villkoren i **MV** med avseende på X och Y, och att X alltså är en orsak till Y (avsnitt 8.7). Detta kräver en granskning av villkoren i **MV** i ljuset av agenter betraktade som biologiska, kausala system i sig själva. Då denna tes därför måste vila på ett antal empiriska antaganden måste den förbli en skiss i monografin.

Om manipulationer tillräckligt ofta är interventioner, och sådana situationer tenderar att sammanfalla med en *upplevelse* av orsakande så kan, givet rätt sorts teori om hur kunskaper kan förvärvas genom direkt upplevelse, dessa upplevelser ge oss direkt, icke-inferentiell kunskap om vissa orsakssamband. Jag föreslår att dessa sinnesupplevelser är de som kallas *proprioception* och *sense of agency*. Jag antar också en naturalistisk, reliabilistisk kunskapsteori för fall av icke-inferentiell, erfarenhetsbaserad kunskap. I korthet ger en typ av sinnesupplevelse U en individ S kunskap om en typ av förhållande P vid ett visst tillfälle om och endast om ett P-förhållande föreligger vid tillfället, U orsakar Ss tro att ett P-förhållande föreligger och ett P-förhållande tenderar att föreligga när S upplever U. (Det vill säga U är tillförlitlig i detta avseende.)

Proprioception är det sinne som säger oss vilka krafter som verkar på vår kropp, samt deras riktning och magnitud (Wolff and Shepard 2013). Proprioception kan därför förmedla information om våra manipulationers direkta konsekvenser (där detta förstås som något fysiskt direkt, som involverar kroppskontakt). Jag hävdar att i kraft av detta kan vi genom en direkt upplevelse veta att villkor 1 och 2 i **MV** är uppfyllda. Villkor 3 i **MV** kräver en mer komplicerad förklaring, som beror av antagandet att kausala effekter minskar ("klingar av") över tid i ett fysiskt realistiskt kausalt system, samt att en agents kognitiva system, delvis på grund därav, är sådant att det är osannolikt att ett starkt probabilistiskt samband mellan en manipulations direkta konsekvens och ett visst observerat utfall förklaras av en gemensam orsak. Ett sådant samband är också en utlösande faktor för vår *sense of agency* (Haggard and Chambon

2012; Saito et al. 2015). Sense of agency är upplevelsen av att vara agenten bakom någon händelse. Om vår sense of agency sammanfaller tillförlitligt (om än inte felfritt) med situationer då ett starkt samband mellan en manipulations direkta konsekvens och ett visst utfall inte förklaras av en gemensam orsak, då kan, givet att villkor 1 och 2 i **MV** också är uppfyllda, instanser av denna upplevelse vara *veridiska*, och ge oss kausal kunskap.

Jag menar alltså att **ICI** kan bidra på detta sätt till en förståelse av hur vi kan förvärva erfarenhetsbaserad, icke-inferentiell kunskap om vissa orsakssamband, på ett sätt som är knutet till våra manipulationer. Denna idé är inte avhängig ofelbarhet i processen—det som krävs är att sense of agency är tillräckligt tillförlitlig för att göra epistemisk nytta. Detta är i kontrast med interventionsteorin avsedd som en tolkning av orsakspåståenden. Men idén är avhängig ett antagande om en reell orsaksrelation, sådan att $\mathbf{T^M}$ är en giltig inferensregel när manipulationen uppfyller villkoren för en intervention.

Slutsatser. Om definitionsteori ger oss villkoren under vilka vi framgångsrikt i en teori angett meningsinnehållet hos ett uttryck, då kan manipulations- och interventionsteorierna som studerats här inte lyckas i det avseendet. Dessa teoriers villkor är antingen för svaga för att vara deskriptivt adekvata, för att de inte utesluter fall där manipulationen har en gemensam orsak med utfallet eller är en gemensam orsak till sin direkta konsekvens och till utfallet, eller så är villkoren istället för svaga för att implicera en välbestämd orsaksrelation givet alla relevanta icke-kausala fakta, på grund av oundvikliga cirkulariteter i teorin.

Men en delmängd av en interventionistisk teori kan, tillsammans med ett primitivt antagande om en reell orsaksrelation som satisfierar denna delmängd, bidra substantiellt till vår förståelse av manipulationers särskilda roll i kausalitetens epistemologi. I en sådan förklaring har manipulation en epistemisk betydelse specifikt för agenten som utför manipulationen. Det står i kontrast med traditionella, regularistiska, förklaringar av manipulationers och experiments roll. I dessa förklaringar gör manipulationer en epistemisk skillnad endast till den grad vi under manipulationen *observerar* något som vi annars inte hade observerat. Vem som utför manipulationen, och vem som istället passivt observerar dess resultat, har alltså ingen betydelse i den bilden. Under den interventionsorienterade förklaringen till manipulationers roll är istället vår epistemiska situation väsentligt annorlunda när det vi observerar är ett resultat av vår egen manipulation, även om vi i övrigt ser precis samma sak som under en passiv observation. Detta förhållande fångas på ett precist sätt av uttrycket "invariance under intervention", som summerar ett tillräckligt villkor för kausalitet inom en del av den kausala inferenslitteraturen.

En interventionsorienterad förklaring av manipulationers epistemiska roll i vårt förvärvande av kausal information bekräftar därmed en intuition som är viktig för manipulationisterna, nämligen den att vår förståelse av kausalitet måste involvera det faktum att vi inte bara är passiva observatörer, utan själva delaktiga som agenter i den kausala världen. Ett annat viktigt motiv bakom en manipulationistisk syn på kausalitet är människors vana att tänka på kausalitet i termer av vad som händer, eller skulle hända, om någon faktor manipulerades.

Detta psykologiska faktum (givet att det är ett faktum) tycks också kunna få sin förklaring i termer av hur vi kan få direkt, erfarenhetsgrundad kausal kunskap på ett särskilt sätt när vi utför manipulationer. Men den förklaringen är avhängig av att kausalitet, eller orsakspåståenden, inte i sin tur ska förstås i termer av manipulation eller intervention.

Bibliography

Anscombe, G. E. M. (1971). Causality And Determination. In Sosa, E., editor, *Causation and Conditionals*. Oxford University Press.

Armstrong, D. M. (1962). *Bodily Sensations*. Routledge and Kegan Paul.

Armstrong, D. M. (1983). *What is a Law of Nature?* Cambridge University Press.

Armstrong, D. M. (1993). Reply to Menzies. In Bacon et al. (1993).

Arntzenius, F. (1992). The Common Cause Principle. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2(227-237).

Arntzenius, F. (2010). Reichenbach's Common Cause Principle. Online. The Stanford Encyclopedia of Philosophy (Fall 2010 Edition): http://plato.stanford.edu/archives/fall2010/entries/physics-Rpcc/.

Ayer, A. J. (1956). What is a law of nature? Norton & Co., 1st edition.

Bacon, J., Campbell, K., and Reinhardt, L., editors (1993). *Ontology, Causality and Mind: Essays in Honour of D. M. Armstrong*. Cambridge University Press.

Baumgartner, M. (2009). Interdefining Causation and Intervention. *Dialectica*, 63(2):175–194.

Belnap, N. (1993). On Rigorous Definitions. *Philosophical Studies*, 72:115–146.

Blackburn, S. (1984). *Spreading the Word*. Oxford University Press.

Blalock, H. M. (1964). *Causal inferences in nonexperimental research*. Chapel Hill:University of North Carolina Press.

Broadbent, A., Vandenbroucke, J. P., and Pearce, N. (2016). Response: Formalism or pluralism? A reply to commentaries on 'Causality and causal inference in epidemiology'. *International Journal of Epidemiology*, 45(6):1841–1851.

Buchdahl, G. (1969). *Metaphysics and the Philosophy of Science*. Basil Blackwell.

Carnap, R. (1962). *Logical Foundations of Probability*. University of Chicago Press.

Cartwright, N. (1979). Causal laws and effective strategies. *Noûs*, 13(4).

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press.

Cartwright, N. (2000). An empiricist defence of singular causes. *Royal Institute of Philosophy Supplement*, 46:47–58.

Cartwright, N. (2007). *Hunting Causes and Using Them*. Cambridge University Press.

Chisholm, R. M. (1966). Freedom and action. Random House.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, 18:115–126.

Collingwood, R. G. (1938). On the so-called idea of causation. *Proceedings of the Aristotelian Society, New Series*, 38.

Collingwood, R. G. (1940). *An Essay on Metaphysics*. Oxford University Press.

Craig, W. (1956). Review of On Padoa's Method in the Theory of Definition by E. W. Beth. *The Journal of Symbolic Logic*, 21(2):194–195.

Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60.

Davidson, D. (1967). Causal Relations. *Journal of Philosophy*, LXIV(21):691–702.

Davidson, D. (2001). *Essays on Actions and Events*. Clarendon Press.

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

Dennett, D. (1986). *Content and Consciousness*. Taylor and Francis.

Donagan, A. (1989). Von Wright on causation, intention, and action. In Schilpp and Hahn (1989).

D'Oro, G. and Connelly, J. (2015). Robin George Collingwood. *The Stanford Encyclopedia of Philosophy (Summer 2015 Edition)*.

Dowe, P. (1992). Wesley Salmon's Process Theory of Causality and the Conserved Quantity Theory. *Philosophy of Science*, 59(2).

Ducasse, C. J. (1926). On the nature and the observability of the causal relation. *Journal of Philosophy*, 23.

Duncan, O. D. (1975). *Introduction to Structural Equation Models*. Academic Press.

Eberhardt, F. and Scheines, R. (2007). Interventions and Causal Inference. *Philosophy of Science*, 74:981–995.

Eells, E. (1991). *Probabilistic Causality*. Cambridge University Press.

Elwert, F. (2013). Graphical causal models. In Morgan, S. L., editor, *Handbook of Causal Analysis for Social Research*. Springer Science.

Floistad, G., editor (1982). *Philosophy of Action*, volume 3. Martinus Nijhoff Publishers.

Freedman, D. A. (2005). Linear Statistical Models for Causation: A Critical Review. In Everitt, B. and Howell, D., editors, *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd.

Freundlich, Y. (1977). The causation recipe. *Dialogue*, 16:472–484.

Fridland, E. (2011). The case for proprioception. *Phenomenology and the Cognitive Sciences*, 10:521–540.

Gasking, D. (1955). Causation and Recipes. *Mind*, 64(256):479–487.

Gijsbers, V. and de Bruin, L. (2014). How agency can solve interventionism's problem of circularity. *Synthese*, 191(8):1775–1791.

Glymour, C. (2004). Review of Making Things Happen by James Woodward. *British Journal for the Philosophy of Science*, 55:779–790.

Glymour, C. (2006). *Markov Properties and Quantum Experiments*. Springer.

Haavelmo, T. (1943). The statistical implications of a system of simultaneous equations. *Econometrica*, 11:1–12.

Haggard, P. and Chambon, V. (2012). Sense of agency. *Current Biology*, 22(10):R390–R392.

Hall, N. (2007). Structural equations and causation. *Philosophical Studies*, 132:109–136.

Hausman, D. M. (1986). Causation and Experimentation. *American Philosophical Quarterly*, 23(2):143–154.

Hausman, D. M. (1997). Causation, agency, and independence. *Philosophy of Science*, 64. Supplement. Proceedings of the 1996 Biennial Meetings of the Philosophy of Science Association.

Hausman, D. M. (2008). *Causal Asymmetries*. Cambridge University Press.

Hausman, D. M. and Woodward, J. (1999). Independence, Invariance, and the Causal Markov Condition. *British Journal of the Philosophy of Science*, 50:521–583.

Hempel, C. G. (1942). The Function of General Laws in History. *The Journal of Philosophy*, 39(2):35–48.

Hempel, C. G. and Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy of Science*, 15(2):135–175.

Hernán, M. A. (2016). Does water kill? A call for less casual causal inferences. *Annals of Epidemiology*, 26(10):674–680.

Hiddleston, E. (2005). Review of Making Things Happen by james woodward. *The Philosophical Review*, 114(4).

Hitchcock, C. (2001a). The Intransitivity of Causation Revealed in Equations and Graphs. *The Journal of Philosophy*, 98(6):273–299.

Hitchcock, C. (2001b). A tale of two effects. *Philosophical Review*, 110(3):361–396.

Hitchcock, C. (2007). What Russell got Right.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association*, 81(396):945–960.

Holland, P. W. and Rubin, D. B. (1983). On Lord's Paradox. In Wainer, H. and Messick, S., editors, *Principals of Modern Psychological Measurement*. Lawrence Erlbaum.

Hoover, K. (1988). *The New Classical Macroeconomics*. Basil Blackwell.

Howard, I. P. (2012). *Perceiving in Depth*. Oxford University Press.

Hox, J. J. and Bechger, T. M. (1998). An Introduction to Structural Equation Modeling. *Family Science Review*, 11:354–373.

Hume, D. (1888). *A Treatise of Human Nature*. Clarendon Press.

Hume, D. (1902). *An Enquiry concerning Human Understanding*. Clarendon Press, 2nd edition.

Humphreys, P., Sober, E., and Woodward, J. (2006). Invariance, Explanation, and Understanding. *Metascience*, 15:39–66.

Kant, I. (1998). *Critique of Pure Reason*. Cambridge University Press.

Kim, J. (1971). Causes and Events: Mackie on Causation. *Journal of Philosophy*, 68:426–41.

Lewis, D. (1970). How to define theoretical terms. *The Journal of Philosophy*, 67(13):427–446.

Lewis, D. (1973). *Counterfactuals*. Basil Blackwell.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4):343–377.
Lewis, D. (1987a). Causal Explanation. In Lewis (1987d).
Lewis, D. (1987b). Causation. In Lewis (1987d).
Lewis, D. (1987c). Counterfactual Dependence and Time's Arrow. In Lewis (1987d).
Lewis, D. (1987d). *Philosophical Papers, volume II*, volume II. Oxford University Press.
Lewis, D. (1989). Dispositional Theories of Value. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 63.
Lewis, D. (1994). Humean Supervenience Debugged. *Mind*, 103(412):473–490.
Mackie, J. L. (1965). Causes and Conditionals. *American Philosophical Quarterly*, 2.4:245–55, 261–4.
Mackie, J. L. (1976). Review of Causality and Determinism by G. H. von Wright. *Journal of Philosophy*, 73:213–218.
Mackie, J. L. (1980). *The Cement of the Universe*. Oxford University Press.
McCain, K. (2015). Interventionism Defended. *Logos & Episteme*, VI(1):61–73.
McCann, E. (1978). Review of Causality and Determinsm by G. H. von Wright. *The Philosophical Review*, 87(1):88–92.
McDermott, M. (1995). Redundant Causation. *The British Journal for the Philosophy of Science*, 46(4):523–544.
Mellor, D. H. (1995). *The Facts of Causation*. Routledge.
Mellor, D. H. (1998). *Real Time II*. Routledge.
Menzies, P. (1993). Laws of Nature, Modality and Humean Supervenience. In Bacon et al. (1993).
Menzies, P. (2006). Review of Making Things Happen by James Woodward. *Mind*, 115.
Menzies, P. (2007). Causation in context.
Menzies, P. (2011). The role of counterfactual dependence in causal judgements. Oxford University Press.
Menzies, P. and Price, H. (1993). Causation as a Secondary Quality. *British Journal for the Philosophy of Science*, 44:187–203.
Mill, J. S. (1882). *A System of Logic*. Harper & Brothers, 8th edition.
Mumford, S. (2009). Causal Powers and Capacities. In Beebee, H., Hitchcock, C., and Menzies, P., editors, *The Oxford Handbook of Causation*. Oxford University Press.
Mumford, S. and Anjum, R. L. (2011). *Getting Causes from Powers*. Oxford University Press.
Näger, P. M. (2016). The causal problem of entanglement. *Synthese*, 193:1127–1155.
Norton, J. D. (2003). Causation as folk science. *Philosopher's Imprint*, 3(4).
Nozick, R. (1969). Newcomb's Problem and Two Principles of Choice. Springer.
Orcutt, G. H. (1952). Actions, Consequences, and Causal Relations. *The Review of Economics and Statistics*, 34(4).
O'Shaughnessey, B. (1995). Proprioception and the body image. In Bermúdez, J. L., editor, *The Body and the Self*.

Paul, L. A. and Hall, N. (2013). *Causation: A User's Guide*. Oxford University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Publishers, Inc.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press.

Pearl, J. (2012). The Causal Foundations of Structural Equation Modeling. In Hoyle, R. H., editor, *Handbook of Structural Equation Modeling*. Guilford Press.

Pearl, J. (2018). Does Obesity Shorten Life? Or is it the Soda? On Non-manipulable Causes. *Journal of Causal Inference (Forthcoming)*.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics: A Primer*. Wiley.

Pearl, J. and MacKenzie, D. (2018). *The Book of Why*. Basic Books.

Prawitz, D. (1989). Von Wright on the concept of cause. In Schilpp and Hahn (1989).

Price, H. (1991). Agency and Probabilistic Causality. *British Journal for the Philosophy of Science*, 42:157–176.

Price, H. (1992a). Agency and Causal Asymmetry. *Mind*, 101(403):501–520.

Price, H. (1992b). The direction of causation: Ramsey's ultimate contingency. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, 2:253–267.

Price, H. (2007). Causal Perspectivalism. Clarendon Press.

Price, H. (2010). *Naturalism Without Mirrors*. Oxford University Press.

Price, H. (2017). Causation, Intervention, and Agency: Woodward on Menzies and Price. Oxford University Press.

Prior, A. (1967). *Past, Present and Future*. Oxford University Press.

Psillos, S. (2014). Regularities, natural patterns and laws of nature. *Theoria*, 79:9–27.

Putnam, H. (1975). *Philosophical Papers, Volume 2: Mind, Language and Reality*. Cambridge University Press.

Ramsey, F. P. (1978). *Foundations*. Routledge & Kegan Paul.

Reichenbach, H. (1956). *The Direction of Time*. University of Los Angeles Press.

Reid, T. (1788/2010). *Essays on the Active Powers of Man*. University Park: Pennsylvania State University Press.

Rosenberg, A. (1973). Causation and recipes: The mixture as before? *Philosophical Studies*, 24:378–385.

Rubin, D. B. (1974). Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, 66:688–701.

Russell, B. (1912). On the Notion of Cause. *Proceedings of the Aristotelian Society*, 13:1–26.

Saito, N., Takahata, K., Murai, T., and Takahashi, H. (2015). Discrepancy between explicit judgement of agency and implicit feeling of agency. *Consciousness and Cognition*, 37:1–7.

Salmon, W. (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton University Press.

Schaffer, J. (2004). Causes need not be Physically Connected to their Effects: The Case for Negative Causation. Blackwell Publishing Ltd.

Schaffer, J. (2007). Causation and Laws of Nature: Reductionism. Wiley-Blackwell.

Schilpp, P. A. and Hahn, L. E., editors (1989). *The Philosophy of Georg Henrik von Wright*. The Library of Living Philosophers. Open Court.

Skyrms, B. (1980). *Causal necessity: A pragmatic investigation of the necessity of law*. Yale University Press.

Spearman, C. (1904). "General Intelligence," Objectively Determined and Measured. *The American Journal of Psychology*, 15(2):201–292.

Spinoza, B. (1677). *Ethics, Demonstrated in Geometrical Order*.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction, and Search*. MIT Press.

Stoutland, F. (1982). Philosophy of action: Davidson, von Wright, and the debate over causation. In Floistad (1982).

Stoutland, F. (1989). Von Wright's theory of action. In Schilpp and Hahn (1989).

Strawson, P. F. (1992). *Analysis and Metaphysics: An Introduction to Philosophy*. Oxford University Press.

Strevens, M. (2007). Review of Making Things Happen by James Woodward. *Philosophy and Phenomenological Research*, LXXIV(1):233–249.

Strevens, M. (2008). Comments on Woodward, Making Things Happen. *Philosophy and Phenomenological Research*, LXXVII(1).

Suppes, P. (1970). *A Probabilistic Theory of Causality*. North-Holland Publishing Company.

Swartz, N. (1985). *The concept of physical law*. Cambridge University Press, 2nd edition.

Tuomela, R. (1982). Explanation of action. In Floistad (1982).

Van Fraassen, B. C. (1989). *Laws and Symmetry*. Clarendon Press.

VanderWeele, T. J. (2018). On Well-defined Hypothetical Interventions in the Potential Outcomes Framework. *Epidemiology*, 29(4).

Violato, C. and Hecker, K. G. (2007). How to Use Structural Equation Modeling in Medical Education Research: A Brief Guide. *Teaching and Learning in Medicine*, 19:362–371.

von Wright, G. H. (1971). *Explanation and Understanding*. Cornell University Press.

von Wright, G. H. (1973). On the Logic and Epistemology of the Causal Relation. *Studies in Logic and the Foundations of Mathematics*, 74.

von Wright, G. H. (1974). *Causality and Determinism*. Columbia University Press.

von Wright, G. H. (1989). The Philosopher Replies. In Schilpp and Hahn (1989).

von Wright, G. H. (1998). *In the Shadow of Descartes*. Kluwer Academic Publishers.

Weirich, P. (2016). Causal Decision Theory. *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*.

Wolff, P. and Shepard, J. (2013). Causation, Touch, and the Perception of Force. *Psychology of Learning and Motivation*, 58.

Woodward, J. (1996). Explanation, invariance, and intervention. *Philosophy of Science*, 64, Supplement. Proceedings of the 1996 BiennialMeetings of the Philosophy of Science Association. Part II: Symposia Papers:S26–S41.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Woodward, J. (2007). Interventionist Theories of Causation in Psychological Perspective. Oxford University Press.

Woodward, J. (2008). Response to Strevens. *Philosophy and Phenomenological Research*, LXXVII(1).

Woodward, J. (2011). Counterfactuals all the way down? *Metascience*, 20:27–52.

Woodward, J. (2014a). From Handles to Interventions: Commentary on R.G. Collingwood, 'The So-Called Idea of Causation'. *International Journal of Epidemiology*, 43(6).

Woodward, J. (2014b). A Functional Account of Causation; or, A Defense of the Legitimacy of Causal Thinking by Reference to the Only Standard That Matters–Usefulness (as Opposed to Metaphysics or Agreement with Intuitive Judgment). *Philosophy of Science*, 81:691–713.

Woodward, J. (2015a). Interventionism and Causal Exclusion. *Philosophy and Phenomenological Research*, XCI(2).

Woodward, J. (2015b). Methodology, ontology, and interventionism. *Synthese*, 192:3577–3599.

Woodward, J. (2016). Causation and manipulability. *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)*.

Woodward, J. (2017). Interventionism and the Missing Metaphysics. In Slater, M. and Yudell, Z., editors, *Metaphysics and the Philosophy of Science: New Essays*.

Woodward, J. (2018). Causal Cognition: Physical Connections, Proportionality, and the Role of Normative Theory. In Gonzalez, W. J., editor, *Philosophy of Psychology: Causality and Psychological Subject: New Reflections on James Woodward's Contribution*. De Gruyter.

Wright, S. (1921). Correlation and Causation. *Journal of Agricultural Research*, XX(7):557–585.

Index

This monograph examines the role of manipulation in theories of causation, with a particular focus on modern theories that aim to explain the meaning of causal claims in terms of what happens under an *intervention*. Beyond the philosophical theories that are reviewed and assessed, the investigation also connects both to new powerful scientific methods of stating and testing causal models, and questions about how our individual familiarity with the phenomenon of causation is formed when we interact with our environment through manipulations.

Henning Strandin

ISBN 978-91-7797-913-5

Department of Philosophy

Stockholm University