Causation as Manipulability and Temporal Direction

by

Elena Popa

Submitted to

Central European University

Department of Philosophy

In partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Supervisor: Professor Ferenc Huoranszki

Budapest, Hungary

2015

Abstract

This dissertation is an inquiry into some problems related to manipulability as an approach to causation, concerning differences between realist and projectivist versions of manipulability, the relation between manipulability and causal realism, the causal and temporal asymmetries and their connection to manipulability, and the extent to which manipulability is a suitable account of causality in psychological contexts.

Chapter 1 discusses realist and projectivist versions of manipulability, as well as related theories of causation. Concerning the debate over the aims Woodward manipulability approach, I argue for investigating manipulability from a metaphysical perspective. I further focus on arguments for connecting manipulability to causal realism, namely a version of the 'No Miracle' argument and an argument from objectivity. While endorsing the former argument, concerning the latter I argue that Woodward's interventionist approach is more or less on a par with the agency theory by Menzies and Price. Finally, I discuss a potential objection against causal realism by Price.

Chapter 2 discusses the problem of the causal asymmetry from a manipulationist perspective and argues that Woodward's particular version of manipulability does not provide a satisfactory account of the asymmetry. After examining several accounts of the causal asymmetry, I propose connecting the causal asymmetry to temporal direction. As a metaphysical claim, I hold that the temporal features grounding the asymmetry are among the more fundamental constituents of causal relations that ground manipulability. A weaker claim connecting manipulability to temporal direction holds that the understanding of the causal asymmetry springs from the understanding of the temporal direction, which is further used in claims about causation and manipulability. I answer the main objections against such proposal. Finally, I use this view to answer Price's objection discussed in chapter 1.

Chapter 3 examines manipulability from a functional perspective. Through psychological data on causal and diagnostic reasoning, I explain why the asymmetry of causation is a problem is functional context as well. I also single out three features of Woodward's account that can be classified as metaphysical assumptions that limit the extent to which manipulability works in functional contexts.

Chapter 4 exemplifies how the previously identified constraints come into place. First, I argue that in developmental context there are several ways of inferring causally, and temporal cues are unaccounted for from the manipulability perspective. I argue that, if connected to a temporal component, the manipulationist concept of causation could reach the generality level of the geometrical-mechanical one. Secondly, I argue that it is difficult to use Woodward's interventionist counterfactuals in cognitive development cases due to the subjects' difficulty of working with counterfactuals. I argue that the agency concept of causation may be more suitable for these cases. Thirdly, contra the non-backtracking interpretation of counterfactuals inherent to Woodward's theory, I show that there are cases that can be accounted for through causal models that allow backtracking. I argue that if counterfactual dependence is not meant to keep causal relations asymmteric, and the asymmetry is dealt with in a different manner, there should be no problem with accepting true backtracking counterfacutals.

I hereby declare that this dissertation contains no materials accepted for any other degrees in any other institutions and no materials previously written and/or published by another person, except where appropriate acknowledgment is made in the form of bibliographical reference.

Budapest, Elena Popa

01.09.2015

Acknowledgements

I would like to thank my supervisor Ferenc Huoranszki for his support throughout the writing process, and comments and suggestions on various versions of this dissertation, David Lagnado for comments on sections of chapters 1 and 4, reading suggestions concerning the psychological literature, and for making it possible for me to spend a semester at UCL, Phyllis Illari for comments on sections of chapters 1 and 2, for helping me develop the argument about the agency account and the objectivity problem, and for pointing out the connection between the functional account and the Canberra plan, Hanoch Ben-Yami for helpful discussions on the causal theory of time and backwards causation, and Daniel Kodaj for comments on an earlier version of section 1.3.

Contents

Introduction

From a broad perspective, this dissertation is an investigation on the relation between causation, manipulability and temporal direction. There will be two main subject areas on which this dissertation will be focusing. One of them is mainly metaphysically-oriented and concerns the possibility of defending causal realism from the framework of a manipulability approach to causation. The other concerns the usefulness (or 'functional' aspect) of such theory with respect to causal reasoning and causal learning. There are at least two ways in which the two parts are linked. On the one hand, as argued in chapter 3, the metaphysical and functional aspects of investigating causation are not completely independent. On the other hand, on both metaphysical and epistemic contexts there seems to be a close relationship between causation and temporal direction. Starting from this claim, I am suggesting that the manipulationist perspective can be connected with an explanation of the asymmetry of causation in terms of temporal direction.

The first chapter discusses the most important contemporary exponents of manipulability-based approaches to causation: Woodward's interventionist account and the agency account by Menzies and Price, as well as Price's later, perspectivalist account. Since Woodward's theory seems to be the only one supporting both manipulability and causal realism, I will be focusing on his account. Consequently, I will also be looking into other accounts that have a significant influence on Woodward's theory: Lewis's counterfactual theory and Bayesian network approaches to causation. The main difficulty in investigating the relation between manipulability and realism in relation to Woodward's theory is Woodward's denial of metaphysics. Drawing from some ways in which the issue is debated in the literature (e.g. Strevens 2007), and some of Woodward's own arguments I will argue for a more metaphysically oriented view on manipulability and causal realism. I will then go on to look into some reasons why one might choose the objectivist take on manipulability supported by realism over the projectivist, agency-based definition by Menzies and Price.

In the second chapter I will look into what I take to be a very important question that philosophical approaches to causation should answer: what is the source of the causal asymmetry? In this respect I will argue that Price and Weslake provide a satisfactory answer for defenders of projectivism, through the perspective of the agent. However, through using an argument by Mackie, I will show that Woodward's concept of manipulability does not satisfactorily deal with the asymmetry problem. I will thus go on and look into possible ways in which the causal asymmetry can be explained by someone defending a manipulability theory along with causal realism. I will end by presenting my own solution, which involves linking temporal direction to the features that ground facts about manipulability. I will also present a weaker version of this claim, in which manipulability can be connected to temporal direction from the perspective of causal understanding. This later claim will be of use in chapter 4, where I will be discussing experimental work involving causal inference through intervention evidence and temporal cues.

In the third chapter I will be turning towards Woodward's considerations on manipulability as a functional account of causation. I will compare this project with the Canberra plan and then expose some criticism from the perspective of causal inference that Woodward's account is vulnerable to. I will conclude by pointing to some shortcomings of Woodward's version of manipulability as a functional account and I will argue that these shortcomings are, in part, the result of metaphysical assumptions underlying manipulability.

In the fourth chapter I will investigate the relation between causation as manipulability and psychological work on causal reasoning and causal learning. I will point to the different kinds of evidence supporting a difference-making as well as a geometrical-mechanical concept of causation and argue that all this evidence cannot be accounted for strictly from the perspective of Woodward’s version of manipulability. I will also show how including a temporal component into a manipulability approach (‘timely intervention’ as discussed by Lagnado and Sloman) may be of help in extending the applicability of manipulability in psychological contexts.

Before proceeding, there are a few things that I should clarify. One of them concerns metaphysics. While I will not be criticising Woodward’s account solely on the grounds of keeping silent on metaphysical issues, I do believe that some of its issues could be solved through metaphysical foundations (as I will try to show with respect to the asymmetry issue). Furthermore, I do not agree with Woodward concerning where the line between metaphysical and functional aspects is drawn. Thus, a significant part of chapters 3 and 4 will be dedicated to showing how claims related to these two aspects intertwine. I will also provide an argument for this stance and illustrate my claim by using examples of how Woodward’s functional project works when applied to psychological data on causation.

The other thing to clarify is the relation between a fundamentally philosophical inquiry into causation and manipulability and experimental data. My aim for the second part of the dissertation can be spelled out as a philosophy of science goal: is the concept of causation, as defined by Woodward, an accurate account of how causal relations are thought of in psychological contexts? My answer is that it has some serious limitations and problems to answer, some of which could be overcome if a closer look were to be taken at aspects such as the asymmetry of causation or the truth value of counterfactuals, which, to some extent, require another look at metaphysics. While I will be looking at experimental data, some traditionally philosophical issues will come about, such as causal unification versus causal pluralism, which one of two or more concepts of causation should be regarded as more fundamental, or questions about backtracking and the semantics of counterfactuals. The investigation in chapter 4 will thus be an illustration of how philosophical discussions can connect to psychological research. Furthermore, the investigation in the second part of this dissertation can also be placed among recent debates around philosophical issues concerning causation where arguments about the origin of causal concepts/causal reasoning seem to be of relevance (as in, for instance, Waskan 2011, Woodward 2011b, Gijsbers & de Bruin 2014).

This dissertation will, thus, incorporate arguments concerning different versions of manipulability theories and their problems, some metaphilosophical considerations about causation, the relation between the metaphysics of causation and experimental practice, as well as an investigation into empirical data and its compatibility with manipulability approaches.

Chapter 1: Manipulability, related accounts, and causal realism

This chapter is meant to set the ground for the problems that I will be looking into throughout the next chapters, namely how can causation be accounted for in terms of manipulability, how it relates to other accounts and projects, and how it fares under causal realism. The first part of this chapter will be mainly an introduction where I will be describing Woodward's manipulability account, the agency account by Menzies and Price, and some related accounts that will have a bearing in points that I will be making in the upcoming sections and chapters. While the link between causation and manipulability can be traced earlier, I will mostly rely on its contemporary versions: Woodward's manipulability account and the agency theory by Menzies and Price. Sections 1.2 and 1.3 will contain a critical discussion over the metaphysical status of causation as manipulability and over the debate between causal realism and causal projectivism.

Throughout this chapter, as well as throughout some sections of subsequent chapters, some debates and issues will be presented from the perspective of Woodward's account as it also contains some criticism of the earlier accounts (i.e. the agency account by Menzies and Price, or Lewis's counterfactual theory). Another reason for emphasizing Woodward's account, as I will be explaining later on in this chapter, is that Woodward's version of manipulability is tied to a realist interpretation which, at first glance, may seem unusual for an account of causation as manipulability. I believe that this option needs further exploration, especially on metaphysical grounds. A final reason is that Woodward rightfully emphasizes and explains the possibility of connecting a concept of causation as manipulability to experimental work involving causes and causal explanations. Nevertheless, as I will be pointing out, there are problems with Woodward's take on causation, and there are cases that the agency concept of causation may handle better. In the forthcoming chapters I will show

how some of these problems may be solved, not necessarily from the framework of Woodward's theory, but also from a more general picture of causation as manipulability. With respect to the usefulness of causation as manipulability in experimental contexts, I will also emphasize how manipulability need not take specifically Woodward's version.

First, I will be presenting Woodward's version of manipulability and its relations to the agency account as well as other two approaches that have important things in common with Woodward's version of manipualbility, the counterfactual account and the causal modeling views. I will then proceed to discussing a debate between Woodward and Strevens over the aims of Woodward's manipulability theory of causation and conclude by emphasizing the need to have a closer look at the metaphysical foundations for manipulability. Secondly, I will take a critical view at the distinction between epistemic and conceptual levels, causal realism and causal projectivism and the problem of objectivity and how they are reflected in the work of Woodward and Menzies and Price. While, as specified earlier, I will endorse both causal realism and the objectivity of causal relations, unlike the view that seems to result from Woodward's discussion, I am going to argue that these claims should be settled on metaphysical grounds. This will prepare the ground for the discussion of the relation between manipulability, the asymmetry of causation, and temporal order, from chapter 2. I will now go on to make a few clarifications concerning some of the distinctions I will be using.

The first issue to clarify is the distinction between a metaphysical and a conceptual level in the analysis of causation. As I will be pointing out in section 1.3, this distinction seems to be sometimes overlooked in both Menzies and Price as well as in Woodward's work, resulting in confusions. Thus, at the metaphysical level, an inquiry into causation is concerned with what causation is, and with what the features of causal relations are. On a conceptual analysis level the main question to ask is how the concept of causation should be defined.

The next issue to clarify will be what I take to be projectivism and realism about causation. Before doing so, since my interest lies in causation as both a metaphysical as well as a conceptual issue, I should mention the two main questions that lie in the background of this investigation:

1. Can causation be defined through agency/intervention? And if so, how should agency/intervention be described?
2. Does causation defined as agency/intervention rest on some objective features of the world, or is it, at least partly,[1] dependent on a human capacity?

In answering (1), both the Menzies and Price (1993) and Woodward (2003) accounts are illustrations of how causation can be defined via manipulability. Their different stances on manipulability lead to divergent answers to the latter question. On the Menzies-Price account, causation is described as a secondary quality, whose instantiation is a matter of a relation between human agents and the world. On Woodward's account, the definition of intervention includes the possibility of hypothetical interventions expressed through counterfactuals, thus allowing for an objectivist picture of causal relations.

Moving on to the realism-projectivism distinction now, I will be using the terms in the following way:

1. From a metaphysical perspective, realism presupposes that causal relations exist 'out there' in the world, independently of human beliefs or capacities. From a conceptual viewpoint, presupposing that causation can be analyzed in terms of a more fundamental concept[2], such analysis should employ more fundamental entities or

[1] The partial dependence on the agent's capacities can be explained through the distinction between metaphysical and conceptual levels: while causal relations may be grounded in objective features of the world, that contribute to our concept of causation, the concept of causation itself is always tied to the agent's perspective.

[2] There is also the possibility of taking causation as a primitive. Since I will mostly be interested in accounts that seek to ground causation in manipulability, and thus, presupposing that causation can be analyzed in terms of a more fundamental concept rather than be taken as primitive, I will not investigate this option. However, this issue will come about when discussing some formal approaches that build models through taking causation as a primitive.

processes that ground causal relations and that can be defined independently from the agent's perspective.

2. From a metaphysical viewpoint, projectivism takes the perspective of a free agent to be constitutive of the existence of causal relations. While projectivism may rely on a more or less objective component to causal relations, facts of causation always involve the agent viewpoint. Conceptually speaking, projectivism holds that the agent's perspective is the key element in analyzing the concept of causation. Once again, agency need not be all there is about causation, but the agent's point of view is in some way written into the concept of causation.

Before going any further, it should also be noted that this distinction does not exhaust all the available possibilities. Another notable option would be to go for a fully epistemic/instrumentalist view. On such view, there need not be causal relations as such:[3] causal claims come in as useful tools when dealing with various contexts, such as scientific explanation, or everyday causal reasoning. Since, as already stated, I will be concerned with causation (among other things) as a metaphysical problem, i.e. what causal relations are, and not why or how causal claims come to be useful, I will only consider this option when linked with Bayes nets (as in Williamson) as a comparison to Woodward's metaphysical assumptions.

Another distinction that will be relevant in what follows is that between what I take to be objective and subjective with regard to causation. On a first glance, from a metaphysical point of view, the earlier considerations on realism and projectivism could be translated into objective and subjective takes on the nature of causal relations: agent-independence versus agent-dependence. However, there are different things at stake with this discussion. I will spell them out in section 1.3, in relation to the interventionist and the agency account.

[3] I.e. no metaphysical level; or at least no interest in explaining the metaphysical level.

1.1 Manipulability and related accounts

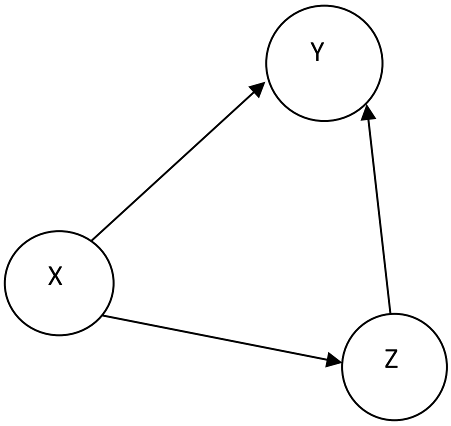
I will now describe how causation can be analyzed through manipulability and how this view relates to other approaches in the causation literature.

1.1.1 Woodward on causation

Woodward argues for a theory of causation and causal explanation based on the ideas of manipulation and control. On Woodward's account, the claim that causation involves manipulation is spelled out in terms of an intervention or a range of interventions on a variable X, leading to changes in another variable Y. Thus, one could say that X causes Y if under an intervention, or a range of interventions to change X, Y would change accordingly. An important thing to point out is that Woodward defines multiple notions of cause, distinguishing between total cause, direct cause and contributing cause. This distinction is employed to deal with some potential objections concerning the necessary and sufficient conditions for a definition of causation. The notion of total cause, amounting also to the simplest formulation of the manipulability account, goes as follows:

> (**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: 51)

Woodward further goes on to stating the necessary and sufficient conditions for the notions of direct cause and contributing cause from the framework of the manipulability theory. To briefly exemplify the distinction, given a case where X causes both Y and Z, and Z at its turn is a cause of Y, one could use certain interventions in order to single out the causal relations.



Thus, if one could intervene on X, while keeping Z fixed, there will be a direct route from X to Y, leading to a change in Y. X would be a direct cause of Y. Along the route X-Z-Y, however, both X and Z would qualify as contributing causes of Y. This characterization, along with his account for direct cause will be sufficient for my purposes here. Woodward defines his notion of direct cause as follows:

> (**DC)** A necessary and sufficient condition for X to be a direct cause of Y with respect to some variable set V is that there be a possible intervention on X that will change Y (or the probability distribution of Y) when all other variables in V besides X and Y are held fixed at some value by other independent interventions. (Woodward 2007: 22)

To sum up, on Woodward's account, causation is a relation between variables and it involves the notion of intervention. The next step is explicating the concept of intervention as a crucial part of the analysis. Woodward is illustrating his concept of intervention by reference to an example when a certain drug's efficacy is tested on a group of patients, contrasting it with a control group, where the drug is not administered. The experimenter has to intervene in order to see if the treatment (T) is causally related to the recovery (R). The first condition involves the notion of a switch. A simple example of a switch is a radio's on/off button.[4] While one could use other buttons to change frequencies or the volume, making a difference in the sound input, these interventions could only work if the radio is on. Thus, the

[4] Example from Woodward & Hitchcock (2003)

on/off button enables or breaks the connections between the other buttons and the sound output. In a more complex setting, a variable could act as a switch if it would break the connection between a variable X and the other variables that could act to change X. Thus, when an intervention acts as a switch, it enables the situation where the value of X is determined exclusively by that specific intervention. This condition is important because it breaks the causal arrows between the variable taken into consideration and the other variables that could change its value. The second condition specifies that if the interventions by the experimenter are correlated with other causes of recovery than T (e.g. placebo), then the reliability of the experiment is undermined. Thirdly, the intervention should not affect recovery independently from T, but, if at all, through it.

More formally put, Woodward's account for an intervention variable goes as follows:

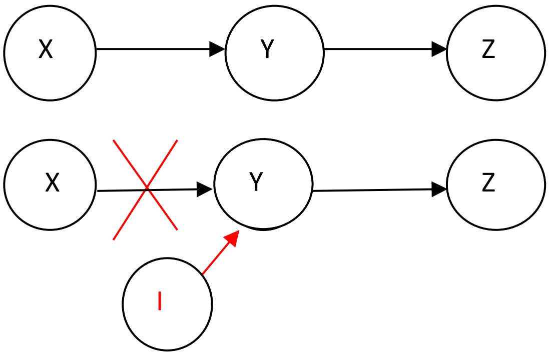
> (**M1**) *I* must be the only cause of *X*; i.e. (…) the intervention must completely disrupt the causal relationship between *X* and its previous causes so that the value of *X* is set entirely by *I*,
> (**M2**) *I* must not directly cause *Y* via a route that does not go through *X* (…),
> (**M3**) *I* should not itself be caused by any cause that affects *Y* via a route that does not go through *X*, and
> (**M4**) *I* leaves the values taken by any causes of *Y* except those that are on the directed path from *I* to *X* to *Y* (should this exist) unchanged. (Woodward 2008[5])

One thing to note is that the arrow-breaking feature of intervention is a common point between Woodward's theory, Pearl's (2000) approach, and the causal Bayes nets literature. One of the main differences is that the latter are more about causal inference, as it is mostly used in statistics and computation while Woodward is willing to turn these ideas into a philosophical account of causation.[6] To show how arrow-breaking interventions work, consider the following graph, where arrows are indicative of causal connections between variables and where assigning a certain value to Y (for simplicity's sake, the values that the variables take can be 1 or 0) amounts to cutting the connection between Y and its previous cause or causes (in this case, X) and seeing what happens to the variables that are taken to be

[5] For a more detailed account, see also Woodward 2003: 98.

[6] I will go into more detail about the causal Bayes net approaches in section 1.1.4.

causally connected to Y. For now, the main point that I will rely on is that Woodward's theory is compatible with such approaches, but struggles for a more philosophical take on some issues concerning causation.



As the previous considerations show, one consequence of this characterization of an intervention is that Woodward's definition of causation is circular, since the concept of cause is used in defining intervention. In Woodward's view, however, this does not lead to a vicious circularity: in order to specify whether there is a causal relation between X and Y one does not need to use the same causal information, but information about I causing X. A further point by Woodward is that even if his account does not aim at reduction, it is not vacuous, leading to different results from other accounts on causation (e.g. causation by omission is possible).

Another important point to note is that in Woodward's version, the manipulability theory requires that a possible intervention or a range of possible interventions on the cause variable would lead to changes in the effect variable, and actual interventions involving human agents are not necessary. In other words, one needs a counterfactual to specify the relation between cause and effect under certain interventions. As Woodward puts it,

> *commitment to a manipulability theory leads unavoidably to the use of counterfactuals concerning what would happen under conditions that may involve violations of physical law*. The reason for this is simply that any plausible version of a manipulability theory must rely on something like the notion of an intervention, and it may be that, for some causal

> claims, there are no physically possible processes that are sufficiently fine-grained or surgical to qualify as interventions. (Woodward 2003: 132-133)

This feature makes Woodward's theory a counterfactual theory. Both features will be important for my discussion in the forthcoming chapters, and I will be coming back to them.

Before going on to pointing out some important implications of Woodward's theory from a broader viewpoint, including related theories on the problem of causation, I will briefly go through Woodward's view on explanation. I will do so mainly to illustrate how the main characteristics of Woodward's take on causation are transferred to his theory of explanation. As I will be pointing out later on, some remarks made in the context of Woodward's counterfactual theory of explanation will be relevant for disclosing some features of the manipulationist concept of causation and its applications to other domains of inquiry.

Similarly to the manipulability account of causation, counterfactuals play an important role in Woodward's theory of explanation. Woodward takes explanation to disclose patterns of counterfactual dependence, or to answer a *what-if-things-have-been-different* question: 'explanation is a matter of exhibiting systematic patterns of counterfactual dependence. Not only can the generalizations (…) be used to show that the explananda (…) were to be expected, given the initial and boundary conditions that actually obtained, but they also can be used to show how these explananda would *change* if these initial and boundary conditions had changed in various ways.' (Woodward 2003: 205)

This commits Woodward's theory to what Salmon (1984) takes to be ontic theories of scientific explanation: causal explanation has the basis in the dependency relations in the world. Finally, another point that I am going to focus on resides in one of Woodward's motivations for his theory of explanation, namely its relation to causation and counterfactuals. As Woodward puts it, the manipulability theory he develops captures the notion of a causal relationship ‘broadly construed, and also shows why the notions of causation and explanation are so closely intertwined: causal claims are explanatory in virtue of providing the sort of counterfactual information that is at the heart of successful explanation.’ (Woodward 2003: 205)

There are a few important things to be mentioned here. First, Woodward assumes that all explanation must fit patterns of causal dependence. This means that his account does not apply in cases where the explanations may not cite causes or in those cases where the explanatory model involves some sort of derivation. Secondly, as it will become more important later on, Woodward explains the asymmetry of explanatory relations through the asymmetry of causation. At a first glance, this seems to be the asymmetry of the manipulability relation: one can manipulate effects through causes, but not the other way around. While I find this account plausible enough when discussing explanatory relations, I believe that the problem of asymmetry can yield into serious difficulties when talking about the dependence relations in the world.[7]

Having presented Woodward’s theory, I will now proceed to a critical survey of some theories on causation that share some of the features of the manipulability theory. In the case of counterfactual and agency theories, I will also show how some of the objections that Woodward raises will be of help in arguing that his view builds up a tension between aiming at articulating a philosophical theory of causation and making no commitment concerning some metaphysical issues about causation.

[7] Briefly put, if Woodward is talking about objective relations in the world, then he has to keep the agent out of the picture. From the way Woodward’s theory is articulated, the obvious place to look would be either his objective interpretation of manipulability, or counterfactual dependence. I will discuss this issue at length in the next chapter.

1.2.2 Woodward's manipulability theory and the agency account

The agency approach by Menzies and Price (1993) is an earlier attempt to articulate a full theory of causation as manipulability. The concept of agency may appear as an influence on Woodward's account: although it shares little of Woodward's formal apparatus of causal graphs, it does link causation to the human capacity of bringing about an effect event through a cause event. The details underlying this approach go in a different direction from the tenets of Woodward's theory. First, Menzies and Price assume a metaphysical stance[8], and the feature that ultimately characterizes causation is related to probability (agent probabilities).[9] Secondly, their account is meant to be non-circular, thus they take some effort to show that the concept of agency is not a causal concept. Third, by considering causation a secondary quality, and suggesting that it should be thought of in the same way as colour, they defend an account where causal relations are not fully objective, or at least not as objective as Woodward takes them to be in his own account. Finally, their way of dealing with unmanipulable causes differs from Woodward's in the sense that they do not rely on counterfactuals. I will discuss these aspects in more detail in this section. I will also rely on Woodward's objections to the agency account in order to disclose some important features of both theories.

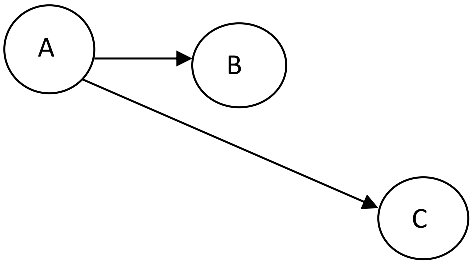
The account proposed by Price's (1991), as well as the one by Menzies and Price (1993) is meant to introduce agency in the analysis of causation in order to deal with the spurious causes and the symmetry objections to probabilistic accounts. The suggestion is that

[8] I take that to be a plausible interpretation of the Menzies and Price original article. In a subsequent article, Price (2014) answers Woodward's criticism by explicitly adhering to what he deems an anthropological (as in non-metaphysical) perspective on causation. While this results in the two accounts being more related than they appear to be from Woodward's discussion, I will stick to the comparison between Woodward's account and the account presented in the Menzies and Price article. The reason for doing so is that I have an interest in metaphysics, and particularly in realism, which seems to be the most important point where Price (2014) disagrees with Woodward. Secondly, as to the metaphysical versus anthropological perspective, I will mostly rely on Woodward's considerations on manipulability as a 'functional' (Woodward 2014) account of causation.

[9] Although I will not be concerned with probabilities here, it is worth mentioning that the agency account can also be viewed as an upgrade to the theories of probabilistic causation, helping in solving some problems they face (see Hitchcock 1993).

probabilities should be taken from the agent’s perspective. Price’s definition of a causal relation between events A and B goes as follows: ‘an event A is a cause of a distinct event B if and only if ensuring that A rather than not-A would be an effective means-end strategy for a free agent whose overriding desire is that it should be the case that B (and whose concern is thus to act so as to maximise the probability that B).’ (Price 1991: 170) Agent probabilities are described as following: ‘agent probabilities are to be thought of as conditional probabilities, assessed from the agent's perspective under the supposition that antecedent condition is realized *ab initio*, as a free act of the agent concerned. Thus the agent probability that one should ascribe to B conditional on A (…) is the probability that B would hold were one to choose to realize A.’ (Menzies & Price 1993: 190)

I will now briefly point out how Price answers the objections to probabilistic approaches by introducing this condition while endorsing the claim that causation is probabilistic. Considering symmetry first, statistical relations are symmetrical, while causal relations are asymmetrical. Contending that causes raise the probability of effects occurring, one has to explain why it does not happen that effects raise the probability of causes. By relying on probabilities connected to the action of a free agent, who uses the causes in order to bring about the effects, only the asymmetrical and not all of the probability relations (namely, including statistical data) are taken into consideration. Secondly, for a spurious causation case let us consider the following situation where A causes B and, at a later time, C. Taking only the statistical dependence into account one could easily get to the conclusion that B causes C (because B is correlated with C). However, on the agency perspective it is impossible to bring about C exclusively through bringing about B. One needs to bring about A in order to see C occurring.



There are several problems that can be raised concerning this approach to causation. I will briefly state them, and present the answers given by Menzies and Price and then move on to Woodward's objections.

The first objection is that the agency theory conflates a metaphysical issue with an epistemological one: while agency may be the most effective way of singling out causes from their effects, it does not need to be built into the analysis of causation. The reply given by Menzies and Price is made by analogy with the dispositional theory of colour: in the same way as colour is defined as relative to being perceived, causal relations are explained by reference to the agents' experience. Thus, 'the concept of causation is to be explained in terms of the way in which an agent's producing, manipulating or "wiggling" one event affects the probability of another event.' (Menzies & Price 1993: 193)

The second objection is mainly a circularity charge: defining causation in terms of events as effective means of bringing about other events already involves a causal concept. Unlike Woodward, Menzies and Price do not want to settle for a circular but non-vacuous account, they go on to arguing that 'bringing about' can be explained through ostension. On the agency account, the concept of causation originates in the early experience of action and bringing about an event in order to achieve another. Thus, 'bringing about' is explained through non-linguistic means, rather than having a causal interpretation.

The third objection concerns unmanipulable causes. The example used by Menzies and Price is that the 1989 San Francisco earthquake was caused by the friction between continental plates. In order to make this causal claim, however, one cannot act upon the continental plates and see whether that action causes an earthquake or not. Unlike Woodward, Menzies and Price reject the counterfactual solution and rely on the principle of analogical reasoning:

> For we would argue that when an agent can bring about one event as a means to bringing about another, this is true in virtue of certain basic intrinsic features of the situation involved, these features being essentially non-causal though not necessarily physical in character. Accordingly, when we are presented with another situation involving a pair of events which resembles the given situation with respect to its intrinsic features, we infer that the pair of events are causally related even though they may not be manipulable. (Menzies & Price 1993: 197)

Thus, cases involving unmanipulable causes can be modelled on cases that agents can bring about, sharing the same intrinsic features (in the earthquake example, the model would be one of the models that seismologists work with).

The final objection holds that the agency account makes causality unavoidably anthropocentric. On this objection, if human beings had different capacities to manipulate the world, then the causal relations would change as well. The reply by Menzies and Price relies on the same principle of analogical reasoning: since unmanipulable causes can be accounted for through this principle, a modification in the degree of human power as agents would at most modify the capacity of understanding potential causal relations. Another point to note from this reply is that while considering causation a secondary quality, Menzies and Price try to show that it is not as dependent on alterations of human capacities and, thus, more objective than qualities such as taste or colour.

A first critique by Woodward refers to how effective agency as defined by Menzies and Price is in resolving spurious causation cases. Taking an example of spurious causation, suppose that a certain virus (V) causes fever (F), and later on red spots on the skin (S). In an experimental setting one could see whether altering F could result in altering S. However, the definition of free action given by Menzies and Price does not exclude the case when one alters F precisely through V and reaches the mistaken conclusion that there is a causal relation between F and S. The problem is that, Woodward points out, in order for an intervention to exclude such cases, one need not be concerned with the S variable, but with the causal connection between F and S, and thus excluding the possibility that V may bring them both about. For Menzies and Price, the problem with such an approach is that it would introduce the concept of cause in the definition of an intervention and render the definition of causation as agency circular.

Another objection by Woodward targets the way in which Menzies and Price deal with unmanipulable causes. Woodward points out that there is no empirically grounded explanation of the step from first person agency to instances where no agent is involved. Even the simple case where, say, in a game of pool an agent hits the cue ball and the cue ball hits the eight ball it is difficult to explain how the former instance of agent causation is transferred to the latter, where there are two balls colliding. This objection mainly points to a distinction made by Woodward (2007) between egocentric, agent-based and fully causal levels of causal understanding. Woodward believes that Menzies's and Price's theory cannot do justice to cases where agency is present (in the case of some animals) but causal knowledge is limited to those actions that the agent can perform.

Apart from the projection problem, Woodward is also sceptical about the efficiency of the solution offered by Menzies and Price. Even though the earthquake can be explained through the agency account by relating to manipulable models used by seismologists, it is hard to explain the resemblance between the two models without using any causal terms. It is actually quite plausible that the resemblances in causal connections be the crucial ones for the isomorphism to hold. Again, this objection targets the idea that the agency account can maintain a non-circular stance in analyzing causation through agent probabilities.

I will not discuss these objections any further, rather, I will take a closer look on some critical considerations made by Woodward with regard to the agency theory and see how they fare when compared with Woodward's overall picture of causal relations.

An important way of delimiting Woodward's theory from the agency account is through the issue of realism about causation. While Menzies and Price take causation to be a secondary quality, a result of the projection of our experience as agents onto the world, Woodward wants to keep an objectivist and realist picture. Woodward opposes the idea that the truth value of causal claims may be dependent on human beliefs, and also the idea that a manipulability approach would necessary involve such a view on causation. His worry is connected to causal relations and the outcomes of controlled experiments which use a manipulability framework: 'it seems very hard to make sense of the activity of conducting experiments to assess the correctness of causal claims if the truth of those claims is somehow partly dependent on or constituted by the experimenter's beliefs or expectations.' (Woodward 2003: 119) Due to his use of counterfactuals in the case of unmanipulable causes, Woodward can say that the counterfactual relations accounting for unmanipulable causes are independent from the human mind. Since I will explicitly argue against this critique later on when discussing realism and projectivism, some further clarifications are necessary.

While according to my reading this is an argument targeting the objectivity with regard to the truth values of causal claims under the projectivist account (and, thus, a conceptual claim), the passage can also be read as a claim about methodology.[10] The latter, methodological, reading would raise the worry that the experiments may yield into results that may be compromised by what has already been projected by the experimenter and would, at the same time, emphasize that interventionism provides a more adequate account for experiments meant to single out causal structures than the agency account, through providing

[10] Woodward, in conversation.

a way of ruling out confounders (i.e. the whole interventionist apparatus). While it is not my purpose here to decide between conceptual and methodological claims about causation and manipulability, I will briefly go on to explain my adherence to the former interpretation. One thing to point out is that Woodward's concept of intervention is better fitted for experimental contexts because it was designed from a methodologically oriented perspective. By contrast, the aims in Menzies and Price are mainly conceptual, or, at most, metaphysical. While on the agency account no method for ruling out confounders is specified, the possibility of such method is not denied either. Explaining the concept of causation through agency does emphasize the role of action as opposed to observation with regard to causal claims, but it leaves the space open with regard to the best experimental methods that reveal causal connections. Thus, if one were to stick to the methodological interpretation, Woodward's objection seems to be merely pointing out that the agency theory does not share interventionism's interest in methodology. Another issue that this interpretation could raise concerns about the elimination of common causes as counfounders. If presented as a counterexample to the agency definition of causation this would constitute a genuine problem for the agency account, and I have previously explained why. However, in the context of the debate between realism and projectivism, Woodward's criticism seems to imply that causal projectivism generally has a problem in dealing with the common cause problem if it takes the objective features of causal relations to be facts about a correlation between the cause and effect variable. Arguably, Menzies and Price are not very clear on what the objective features of causal relations are, but neither is Woodward. I will get back to this issue in section 1.3, when discussing objectivity. Regardless of the way in which this claim may be interpreted, I believe that the problem of the objectivity of causation and of agent dependence in an important part of the issue. Consequently, I will go with the conceptual interpretation while noting that even if one sticks to the methodological interpretation there seems to be a worry regarding causal projectivism, the objectivity of causation, and experimental practice.

Another argument for realism stems from an evolutionary perspective on the human agency and the knowledge of causal structures: while there is no evolutionary story explaining how the projective activity may yield into any benefits, it is plausible to suppose that the ability to distinguish accidental correlations from causal relations is increasing survival rate if such a difference exists objectively and affects survival.[11]

At the same time, though, Woodward is trying to avoid going into a fully metaphysical picture:

> I emphasize that the kind of realism that follows from this way of viewing matters is metaphysically modest and noncommittal. It requires only that there be facts of the matter, independent of facts about human abilities and psychology, about which counterfactual claims about the outcome of hypothetical experiments are true or false and about whether a correlation between *C* and *E* reflects a causal relationship between *C* and *E* or not. (Woodward 2003: 121)

Given that Woodward makes this claim in comparison with other approaches to manipulability, it seems to be unproblematic at this point. However, as I will argue in section 1.2, Woodward is willing to maintain this noncommittal stance even on issues that concern more than the objectivity of nonmanipulable causes. For now, I will only point out that Woodward abides to a realist picture on causation but is not going into the metaphysical details about what facts about causation and facts about manipulability and their relations may be.

One last point to mention, regarding the way in which Woodward, on the one hand, and Menzies and Price, on the other, deal with unmanipulable causes is that Woodward believes that there is an objectivist interpretation to the idea that models involving

[11] Woodward's example is the correlation between eating a certain mushroom and feeling nausea. One could see why in this case the individuals who managed to establish a causal relation would have increased chances of survival than those who did not, and why an objective view on this instance of a causal relation is important.

manipulable variables share their intrinsic properties with real world instances of unmanipulable causes. To use the quote from Woodward at length:

> 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. This view yields a far more plausible treatment of causes that are not manipulable by human agents and avoids the problems that result from taking agency to be a primitive feature of the world, but it also abandons any pretense of noncircular reduction of causality to agency. (Woodward 2003: 125-126)

The important conclusion to draw from here is that the objective picture that Woodward supports does not take manipulability to be a primitive feature of the world, thus providing a nonreductive approach to causation. Also, after reading this passage, someone who is more interested in the metaphysics of causation may ask what causation is according to Woodward if we are to leave the epistemic considerations aside for a moment. I believe that this is far from clear in *Making Things Happen* and I will illustrate the problems that spring from here in section 1.2.

1.1.3 Counterfactuals and manipulability

As specified earlier, Woodward relies on counterfactuals to explain instances of unmanipulable causes and to articulate his account of scientific explanation. A question that arises here is to what extent is Woodward's theory similar to counterfactual approaches, particularly to Lewis's account. A further question concerns the ways in which Woodward's theory departs from this framework. In this section I will discuss Lewis's earlier counterfactual theory and Woodward's take on it. I will then go on to making a few observations about Lewis's later account of causation as influence. The influence account is not discussed by Woodward despite the fact that, as I will argue, it seems to be closer to the tenets of the manipulability view. Some issues concerning counterfactuals introduced here will come up again in the next chapters, notably in relation to backtracking and the asymmetry of causation.

Lewis (1973) analyzes causation through counterfactual dependence. On Lewis's view, X causes Y if X and Y are distinct events and if the following counterfactual is true: 'If X had not occurred, Y would not have occurred.' This accounts for causal dependence. Lewis holds that causation is the ancestral of causal dependence: X causes Y if Y causally depends on X or if there is a chain of causal dependence between the two. According to Lewis's semantics of counterfactuals, taking the actual world where X occurs and Y occurs and then looking at worlds where X does not occur, the worlds in which X does not occur and Y does not occur are closer to the actual world that worlds where X does not occur and Y occurs.

Given this analysis of causation and Lewis's further aims to explain both the direction of time and the direction of causation through counterfactual dependence, counterfactual dependence needs to be an asymmetric relation.[12] Some examples, like the one by Downing (1959), show that backward counterfactuals need an analysis consistent with the considerations on causation and time, if the two are to be explained in terms of counterfactual dependence. The example goes as follows: supposing that Mr. D'Arcy had a fight with Elisabeth and later (1) if Mr. D'Arcy were to ask Elisabeth for a favor, she would refuse him. But given D'Arcy's pride, (2) he would never have asked for a favor if they had had a fight the previous day. Therefore, (3) if D'Arcy would ask for a favor, Elisabeth would grant it. If (2) is true, then both (1) and (3) are true, which leads to a contradiction. Lewis's answer to this problem goes along the line of taking such backtracking counterfactuals to turn out false. He opts for an analysis where some counterfactuals have to be disambiguated:

[12] I am only introducing the asymmetry of causation issue here; I will go into more depth about it in chapter 2.

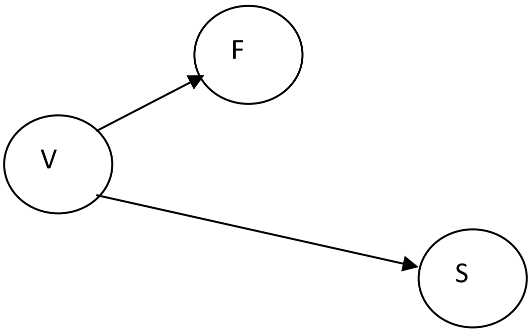
> (1) Counterfactuals are infected with vagueness, as everyone agrees. (…) (2) We ordinarily resolve the vagueness of counterfactuals in such a way that counterfactual dependence is asymmetric (…). Under this standard resolution, back-tracking arguments are mistaken: if the present were different the past would be the same, but the same past causes would fail somehow to cause the same present effects. (Lewis 1979: 457)

As Hausman notes, Lewis has to show that while effects counterfactually depend on their causes, those counterfactuals involving dependence of causes on effects are false: 'in order to account for the asymmetry of causation, Lewis need only deny the counterfactual dependence of specific causes on their effects.' (Hausman 1998: 144) That is to say, his analysis has to rule out the counterfactuals that track the effects back to their causes. Lewis's analysis goes as follows: 'A counterfactual "If it were that A, then it would be that C" is (non-vacuously) true if and only if some (accessible) world where both A and C are true is more similar to our actual world, overall, than is any world where A is true but C is false.' (Lewis 1979: 465)

Lewis further goes on to asserting the criteria of similarity between worlds:

> (S1) It is of the first importance to avoid big, widespread, diverse violations of law.
> (S2) It is of the second importance to maximize the spatio-temporal region throughout which perfect match of particular fact prevails.
> (S3) It is of the third importance to avoid even small, localized simple violations of law.
> (S4) It is of little or no importance to secure approximate similarity of particular fact, even in matters that concern us greatly. (Lewis 1979: 472)

After spelling out Lewis's view, it becomes clearer how Woodward's view relates to Lewis's theory through the use of counterfactuals. I will only focus on the aspects relevant to the points that I will be making. Taking the case of X causing Y, and the counterfactual 'If X had not occurred, Y would not have occurred', the antecedent (X not occurring) is established through a small miracle (in Lewis's approach) or through an intervention (in Woodward's theory). It should be noted that, despite the different apparatus behind Woodward's theory, on both accounts it ultimately comes down to looking into a counterfactual scenario in order to single out causal connections. Another interesting thing to note is that the arrow-breaking feature of interventions can be accounted for from within Lewis's account. In order to render this point clear, let us recall the case of spurious causation discussed above where a virus (V) causes both fever (F) and red spots on the skin (S), with F occurring some time before S, like in the following diagram.



Given that F is always followed by S, one could think that the causal chain goes from F to S. Woodward's solution to this would involve intervening to fix F's value through a different variable, I, while in the same time breaking the causal connection between V and F. In this case, no changes would be observed in S, and one could infer that F is not the cause of S. Lewis's treatment of such a case, going back to his 'Causation' paper would be to consider a possible world where V occurs, but it fails to produce F, while S still occurs. This shows that there is no counterfactual dependence between F and S. While this method is not applicable to a scientific investigation in separating the causes of a disease from its symptoms, from the metaphysical perspective under discussion at this point, it is clear that Lewis's framework could accommodate the arrow-breaking feature of potential interventions. Finally, perhaps the most important similarity for my purposes here is that both Lewis and Woodward rely on a non-backtracking interpretation of counterfactuals. For Lewis, the most obvious motivation seems to be keeping the asymmetric character of causal relations. While the same motivation could apply to the Woodward account, another potential explanation could be linked to Woodward's concept of intervention and his use of causal models.

Turning to the differences side, there are several respects in which the two accounts part ways. A very important one concerns the problem of reduction. Lewis's account is clearly targeting an analysis of causation in terms of counterfactual dependence. That is why he avoids the use of any causal concepts in describing counterfactual dependence and in stating the similarity criteria. Woodward's approach, through the causal definition of intervention, is not aiming for a non-circular analysis. Furthermore, Woodward believes that certain counterexamples that threat Lewis's theory could be answered by dropping the reductive perspective. Secondly, Lewis's motivation for the similarity criteria is exactly to rule out backtracking counterfactuals, and does not correspond to any practical features of causal thinking. Unlike Lewis, Woodward's reliance on manipulability and control provides his theory with a connection to the capacity of intervening in order to single out causally connected variables. Thirdly, Woodward points out that Lewis's similarity criteria are imprecise. It is difficult to tell what a large miracle is, or whether two small miracles would make a large miracle. This objection is especially strong when one would aim at using similarity criteria for evaluating causal claims in science. Setting up a controlled experiment where one finds a way to set the value of a cause variable and see the modifications in the effect variable is much more precise than talking about miracles and possible worlds. A similar point is made in Woodward & Hitchcock 2003 (8-9) where the notion of invariance is compared with Lewis's similarity criteria presented above. Springing from this critique, Woodward and Hitchcock also point out that it is doubtful whether these criteria can be spelled out in purely non-causal terms. Finally, unlike Lewis, Woodward does not assume that causation is a transitive relation.

After discussing these divergences, Woodward goes on to show that there are cases where Lewis's account fails to identify the right counterfactuals and causal structures, whereas the manipulability theory can get to the right solution. I am now briefly going to discuss one of them. Supposing that C1 causes C2, C3… Cn, and, independently from C2-Cn, C1 causes E. On a counterfactual theory of causation, the following counterfactual should turn out false 'If C2, C3, … Cn had not occurred, E would not have occurred.'

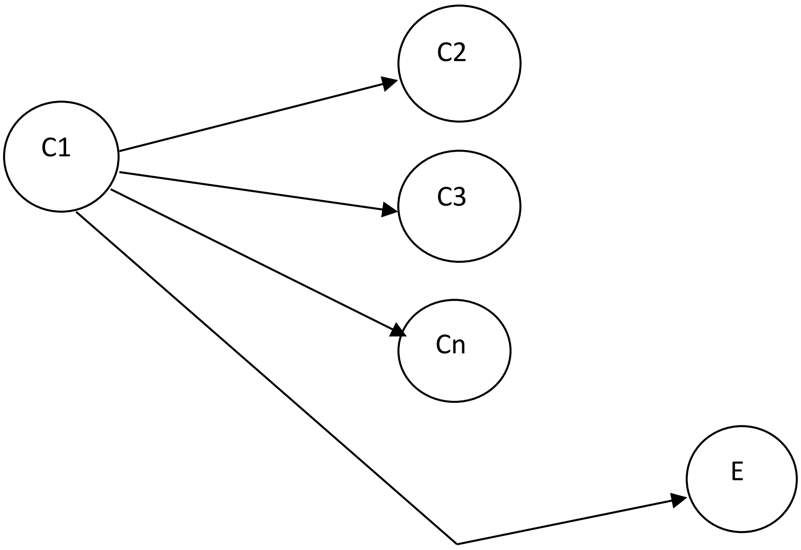


figure taken from Woodward 2003: 139

There are two ways in which Lewis could deal with this problem. One would be to let C1 occur and have n-1 miracles to prevent C2-Cn from happening. That would, however, go against Lewis's first criterion, of avoiding widespread violations of laws. The world closer to the actual world would have to have a miracle before C1, preventing C1 from happening. However, in this world the counterfactual would turn out true, since both C1 and C2-Cn would not occur, preventing E from occurring. Moreover, in such a world the following backtracking counterfactual would be true as well: 'If C2-Cn had not occurred, C1 would not have occurred.'

Intuitively, one would say that the former possible world should be taken into consideration when evaluating the counterfactual. Woodward's concept of intervention allows for this possibility, because interventions can break the arrows from C1 to variables that are not causally relevant to E. Also, because an intervention on C2-Cn should not change other variables than those that stand causally between C2-Cn and E, removing C1 could not be part of the intervention. Both problematic counterfactuals should turn out false.

This counterexample shows not only that Woodward's concept of intervention deals better with problems related to the truth value of counterfactual claims, but also that there are cases that Lewis's account cannot resolve unless some concept of cause is used. Thus, the way in which Woodward defines his concept of intervention is both clearer in terms of applicability in scientific claims, as well as better suited for dealing with counterexamples.

As mentioned before, Woodward does not discuss Lewis's later, influence, account in *Making Things Happen*. I believe, however, that the influence picture of causation is in certain respects very similar to the core ideas behind manipulability. For this reason, I will first describe Lewis's (2004) account, developed in order to answer preemption cases[13] that trouble the counterfactual analysis of causation. Lewis is relying on the concept of an alteration of an event. An event alteration is fragile in the sense that any change in the time or manner in which an event occurs amounts to a different event alteration. Lewis describes influence as follows:

> Where C and E are distinct actual events, let us say that C *influences* E iff there is a substantial range C1;C2; . . . of different not-too-distant alterations of C (including the actual alteration of C) and there is a range E1; E2; . . . of alterations of E, at least some of which differ, such that if C1 had occurred, E1 would have occurred, and if C2 had occurred, E2 would have occurred, and so on. Thus we have a pattern of counterfactual dependence of whether, when, and how on whether, when, and how. (Lewis 2004: 91)

[13] See Shaffer 2000.

Lewis explains causation as related to patterns of influence through describing the world in terms of a complicated machine where each part is connected to another part and 'you want to find out which bits are connected to which others. So you wiggle first one bit and then another, and each time you see what else wiggles.' (Lewis 2004: 91) As Lewis further puts it, knowing which counterfactuals are true amounts to knowing the causal relations in the world, given that one cannot wiggle events. It is interesting to point out that Woodward and Hausman make use of the same metaphor when they speak of manipulability: 'If X causes Y, one can wiggle Y by wiggling X, while when one wiggles Y, X remains unchanged.' (Hausman & Woodward 1999: 533) Another interesting point to note here is that Woodward uses this very feature to explain unmanipulable causes: when actual intervention is impossible, causal claims can be evaluated in terms of knowing which counterfactuals are true.

Further connections can be found in Woodward's take on causation and sensitivity. Unlike invariance[14], that requires changes in the variables' values, sensitivity involves changes in background conditions. One way for Woodward to explain sensitivity is through showing how a relation where a cause has a great influence, in Lewis's sense, on the effect is going to be insensitive, that is, less likely to change under modifications in background conditions.[15] While the apparatus behind manipulability is different from the one behind Lewis's counterfactual theory, the point that I wish to emphasize is that Woodward's and Lewis's picture of the causal relations in the world seem to be to an important extent similar.

One thing worth addressing here is the fact that Woodward considers patterns of counterfactual dependence between variables, not events and alterations of events. This difference, however, is not crucial; patterns of influence between the parts of a machine could be spelled out both in terms of events, alterations and fragility and in terms of variables whose

[14] See, for instance, Woodward 2003:15-16.
[15] See, for instance, Woodward 2006: 30.

values could be modified, leading to changes in some other variables' values. One could intervene to change certain variables, and where intervention is impossible, it comes to knowing which counterfactuals are true. Furthermore, the distinction between events and alterations of events could perfectly match the distinction between a variable and its values. For instance, the variables treatment (T) and recovery (R) can have the values 1 or 0, in Lewis's terminology, setting T to 0 and seeing what happens to R would amount to going from one alteration of an event to the other. This also shows that, as far as the problem of causal relata is concerned, Lewis's account can handle the more complex cases present in science, when variables can take more than binary values. Spelling this problem out in Strevens's (2014) terms, it all seems to come down to the relations between physical constituents of the world, specified as concrete events.

A final point to make about the comparison between the two accounts concerns Lewis's very idea of viewing the causal connections in the world as patterns of influence within a complex machine. This could correspond to one of the main features of Woodward's view on causation: one can intervene and see which variables change only if they are part of a system. Lewis seems to rely on a similar idea when talking about patterns of influence. Also, both views seem to encounter some problems when having to consider the world as a system, as one could not bring about an intervention variable from outside the system.[16]

I should point out that the resemblances I have been discussing are more about the more general, metaphysical picture. While one might argue that the two accounts are alike because they both claim that causation involves making a difference, I would argue further, that, even though Woodward does not explicitly endorse any metaphysical aims, if there were to be some metaphysical foundations to manipulability they would very much resemble the picture sketched in the influence account. It is also important to note that Woodward does not

[16] Woodward contends that no interventions are possible when the whole world is taken as a system, a consequence of this being to admit that the interventionist account does not fit so well the claims of fundamental physics (for more on this, see Woodward 2007b).

mention what seems to be the most important difference: Lewis's aim at offering an account of the metaphysics of causation, while Woodward denies making such an attempt when articulating the manipulability approach. Before addressing this issue, I will have a look into a different set of approaches to causation that relate to Woodward's version of manipulability: probabilistic and causal modeling approaches.

1.1.4 Probabilities, causal modelling, and Woodward's manipulability account

Another way of looking at Woodward's theory of causation would be through analyzing its similarities with probabilistic approaches. While Woodward does not defend a probabilistic account of causation, his theory accommodates a concept of causation that involves a change in probability distribution of the effect-variable (an intervention can change an effect, or the probability distribution of an effect through changing its cause). The use of directed acyclic graphs and the arrow-breaking feature of interventions, bring the manipulability theory close to causal modeling approaches. In this section I will briefly survey the issue of probability and its uses in the problem of causation, focusing on causal Bayes networks and on Pearl's theory. Furthermore, I will show how Woodward's theory relates to causal modeling, and whether this connection imposes any particular features on the manipulability approach.

1.1.4.1 Causation and probability. Interpretations of probability

Before going into the causal modelling approaches, it will be useful to discuss some distinctions and interpretations of probabilities, as well as their connection to the problem of causation. Briefly put, a probabilistic theory of causation is based on the claim that causes change the probability of their effects occurring. While the issue of probability will come about a few times in the chapters to follow, I will not subscribe to a particular version of probabilistic theories.

In order to understand what probabilistic approaches to causation are working with, an overview over the interpretations of probability is necessary. According to Williamson's (2005) classification and discussion, the interpretations of probability can be classified according to the following distinctions:

a) Single-case (token level) vs. repeatable (type level).

b) Mental vs. physical.

c) Subjective vs. objective (if two agents with the same background knowledge can disagree with respect to the probability value and they are both right, then we deal with a subjective interpretation, otherwise it is objective).

These distinctions are useful in describing the following interpretations of probabilities:

1. The frequentist interpretation – the variables are repeatable instances in a set; they are physical, and objective.
2. The propensity interpretation – Popper's attempt to apply the frequentist interpretation to single case variables. It includes the Axiom of Independence, specifying that in the case of collectives resulting from a repeatable experiment probabilities are the same.
3. The chance interpretation – applies to single case variables, based on chance fixers. A chance fixer usually refers to the configuration of the world up until the moment the probability is being ascribed.
4. Bayesianism – deals with both repeatable as well as single case probabilities, but from a mental, rather than physical point of view – probabilities are an agent's degrees of belief. There are both subjective as well as objective interpretations of Bayesianism.

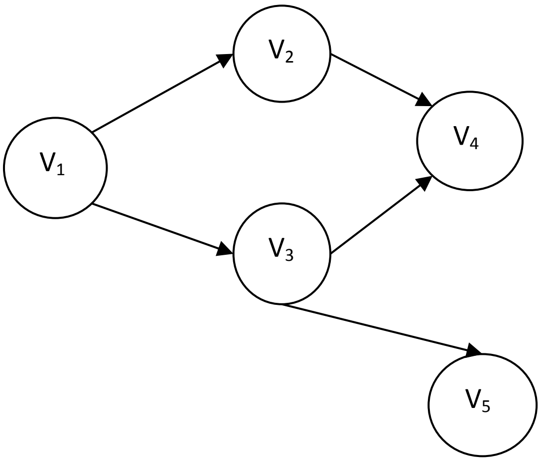
In articulating a probabilistic approach to causation, one can take one of these stances towards probabilities, yielding into theories of causation with different implications. For instance, as we shall see later, Williamson's approach relies of causal Bayes networks, on an objective interpretation of Bayesianism, and on an epistemic view of causality.[17]

1.1.4.2 Bayes networks and causality

A Bayesian network, or, for short, Bayes net is a directed acyclic graph (DAG) *G* containing a set of variables (V), and their probability distributions. The arrows connecting the variables indicate whether there is direct influence from one variable to another. If there is an arrow from variable V_1 to variable V_2, then V_1 is the parent of V_2 (and V_2 is the descendant of V_1). Further, the nodes in the network are annotated with the conditional probability distribution (CPD), for instance, $P(V_1/(ParV_1))$ (the probability of V_1 given the parents of V_1). The pair (*G*, CPD) encodes the joint distribution $p(V_1 \ldots V_n)$. The joint probability distribution over V from G is factorized as:

$P(V_1 ... V_n) = \prod n \ (p(V_n/Par(V_n)))$

An example[18] of a Bayes net would be:



The conditional probabilities are specified for each node in the graph, given the possible values that the variables could take. Just as an example[19], supposing that the possible

[17] Different from an epistemic view on probability.

[18] Both the example and the description are a modified version of an example from www.bayesnets.com.

[19]This is just to illustrate how a Bayesian networks, the probability distributions will have no bearing the subsequent discussion about Bayesian networks and causality.

values of the variables are 0 and 1, the probability distributions for node V4 to take value 0 or 1 conditional on the V2 and V3 nodes could go as follows:

V2	**V3**	**V4=0**	**V4=1**
0	0	0.7	0.1
1	0	0.5	0.2
0	1	0.3	0.5
1	1	0.1	0.8

Further, the graph and the probabilities are held together by the *Markov Condition*. The Markov condition specifies that a variable is conditionally independent from its non-descendants given its parents. For instance, in the presented example, V2 is independent of V3 given V1. That means that p(V2/(V3V1)) = p(V2/(V1)). Other independencies in the graph are: V2 is independent of V5 given V1; V3 is independent of V2 given V1; V4 is independent of V5 and V1 given V2 and V3; V5 is independent of V4, V2 and V1 given V3. It should be noted that the Causal Markov condition is a more general principle, entailing Reichenbach's Common Cause Principle and its screening off relations in a causal context. For instance, if two variables have a common cause, the common cause will screen off its effects from one another. To use an example from the graph, according to the screening off, condition the following relations hold:

p(V2/V1)>p(V2/~V1)

p(V3/V1)>p(V3/~V1)

p(V2&V3/V1)= p(V2/V1)*p(V3/V1)

p(V2&V3/~V1)=p(V2/~V1)*p(V3/~V1).

What is interesting to note is that the probabilities in the graph need not have a Bayesian interpretation. As Williamson specifies in a footnote, what is 'Bayesian' in a Bayesian net is the fact that it is updated through Bayesian conditionalisation (the conditional probabilities specified in each node);[20] the probabilities can have an interpretation different from the Bayesian one. (Williamson 2005: 51)

In the application of Bayes nets to causal relations, the arrows connect causally relevant variables. In a graph where two variables X with possible values x1 and x2, and Y with y1 and y2 as possible values are causally connected, one could change the value of Y by changing the value of X. Using Pearl's (2000) *do* operator (which can be read as a intervening to set a variable's value)[21] the causal relation would be written as: $p(Y/do(X=x1)) \neq p(Y/do(X=x2))$. As Williamson specifies, Bayes nets make three assumptions about causal relations: (1) that they are direct relations between variables; (2) that they are acyclic; (3) that they respect the Causal Markov condition.

According to Hitchcock (2011), the causal modelling approach, which uses Bayes nets, is used to solve two kinds of problems concerning causality. The first concerns discovering the qualitative causal structure through information about probability correlations and previous hypotheses about the causal structure. The second is about the identification of causally significant quantities (such as the probability of an intervention producing a certain effect) through observed probabilistic correlations and qualitative causal information. Thus, the Bayes nets constitute a good tool for inferring causal structures from information about probabilities and also for extracting quantitative[22] information from previously known causal structures.

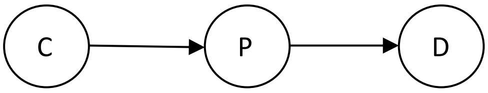
I will now proceed to showing how this formal apparatus would work on a causal structure. Let us suppose that in a province of an imaginary country there is a worry about people's increasing discontent about the government. A closer look at the province shows both high levels of crime and high levels of poverty. Taking the relevant variables to be crime

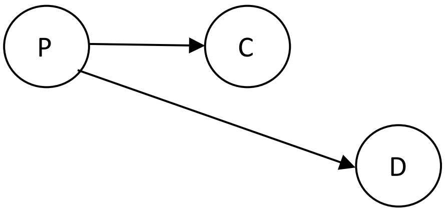
[20] The Bayesian view on evidence, where the degree of belief in a certain hypothesis is updated through the probability of the hypothesis given the evidence, is another instance where conditional probability is at work.

[21] This will be explained further in the next subsection.

[22] While Hitchcock mentions only quantitative information, qualitative information could fit in as well. I am only mentioning this for clarity's sake, as this issue will not be important for my discussion here.

(C), poverty (P), and government discontent (D), let us suppose that there are two causal structures that could be used to account for the case. To keep it simple, I will only use a set of values without going into the probability distributions, it should be noted, though, that the relations are probabilistic. The graph representations go as follows:





An important thing to note is that sticking to the conditional probability approach, based on observation is of no help in finding out the right structure. On both structures, C and D are conditionally independent given P. Given that all three variables are already present in the current configuration, causal judgment based on observation is not going to be very informative either. In a model where intervention is possible, however, the two causal structures can be differentiated. Assuming that the common cause structure holds, one could intervene on C without altering P (say, by sending more police officers in that province) and see whether there are any changes in D. In this particular example, under a common cause structure changing C would not bring about changes in D, while on a causal chain scenario, changing C would bring about changes in D through P. Thus, the strength of the Bayes net approach is apparent in the way one can learn the causal structures from the graph and in terms of the possible values of variables.

I can now point out how Woodward's theory relates to the causal Bayes nets formal approach. First, his definition of causation uses a concept of intervention compatible with the constraints of the Bayes nets approach. Establishing the causal connections between P and C and D in the previous example can very well be spelled out in terms of the interventionist notion of causation. Secondly, he assumes that causation is a direct relation between variables, and that causal relations could be captured in an acyclic graph, which is the case for the causal Bayes nets as well. Thirdly, just as in the case of the formal approaches the system must satisfy the Causal Markov condition. This may bring about some objections, which I will not be discussing (see Cartwright 2002 for criticism, and Hausman & Woodward 1999 for a defense).

Having given a brief introduction to Bayes nets in this section, I will move on to some of the considerations of Pearl and Williamson on causality both of whom rely on Bayes nets in their theories of causation. I will present Pearl's ideas as a common ground with Woodward and see how Pearl's computational approach lends its advantages to Woodward's philosophical considerations. On the other hand, Williamson's theory will be a good example of the employment of probabilities and causal Bayes nets that yields into an epistemic view on causation.

1.1.4.3 Probability theory, causality, and intervention

I will now highlight the common points between Pearl and Woodward's approaches, along with the main differences. My purpose will be to stress the common assumptions, as well as the formal apparatus that Woodward's manipulability theory shares with Pearl's work on causal inference. An important thing to emphasize is that in the chapters to follow I will only focus on Woodward's theory, and its philosophical underpinnings. Thus, I will briefly describe Pearl's approach only with the purpose of clarifying the background for Woodward's manipulability account.

Pearl’s (2000) approach starts from the premise that the language of probability calculus cannot capture causal relations, and that these relations cannot be inferred from statistical data alone. Some of his examples of statistical concepts are ‘correlation, regression, dependence, conditional independence, association, likelihood, collapsibility, risk ratio’ (Pearl 2001: 29), whereas causal concepts would be ‘randomization, influence, effect, confounding, disturbance, spurious correlation, instrumental variables, intervention, explanation, attribution’ (Pearl 2001: 29). According to Pearl, the problem with causal claims, from the perspective of probability theory and Bayesianism is that they cannot be verified unless one resorts to experimental control.[23] His solution to the problem of causation and probability theory includes three steps:

1. Taking causation as a summary of behaviour under interventions.
2. Using equations and graphs as a mathematical language which allows causal claims to be represented and manipulated.
3. Putting these two steps together with a concept of intervention as surgery over equations. (Pearl 2000: 344)

Looking at the basis of Woodward’s approach, one could see (1) causation involving invariance under a range of interventions; (2) representing causal relations and interventions through DAG-s; and (3) the definition of interventions through their arrow-breaking feature. Taking these steps together shows that both Pearl and Woodward take precisely experimental control to be relevant in explaining the nature of causal relations. Furthermore, Woodward’s reliance on counterfactuals can be traced back to Pearl’s idea that what he takes ‘deep understanding’ to involve is not only knowing how things work, but also how they would work on certain interventions: ‘deliberate reasoners (…) can anticipate the consequences of new manipulations without ever trying those manipulations. (Pearl 2000: 346) This issue is

[23] This does not seem to be always true, however; observational evidence can sometimes provide satisfactory evidence for causal claims.

controversial in the probability and causal modelling literature, as well as in the philosophical literature. I will not go into the technical difficulties surrounding Pearl's use of counterfactuals.[24] Instead, in the next chapter, I will point out how some problems of counterfactual theories of causation, most importantly, the backtracking issue, affect Woodward's theory as well.

Another reason why probability theory cannot accommodate causation is that it focuses on observation. Classical talk of conditional probabilities takes the form of, say, variable Y given variable X (Y/X), which reads as the chance of Y occurring given that X is observed. Pearl's suggestion is that the probability calculus should also include the situation when one investigates over the chance of Y occurring given that one does X. An example of this would be evaluating the probability of an individual having the common cold given that he has fever. Making a causal claim on the bases of observational evidence could be lead to the unpleasant outcome of not making the difference between the disease an its symptoms: since they are always observed together, one could just as well say that it is fever that causes the common cold. Upon using an intervention, say, administering a medicine that stops the fever, and still observing the other symptoms of the common cold, one could single out the common cold as the cause of the fever, and not the other way around.

Pearl's way of dealing with this problem is through building a mathematical approach to probability and causation that includes the *do* operator. The *do* operator is part of an axiomatic system developed in order to determine the results of interventions. Pearl uses three axioms that specify how the *do* operator can be substituted by relations used in the probability calculus.[25] Finally, Pearl's definition of cause goes as follows: 'Given (…) two mathematical objects (…) variable X is a *probabilistic-cause* of variable Y if P(y/do(x)) ≠ P(y) for some

[24] See Pearl 2011 for this.

[25] The axioms are: (1) Ignoring irrelevant observations; (2) Action/Observation exchange for the same facts; (3) Ignoring irrelevant actions. I will not go into detail about them.

values X and Y. Since each of P(y/do(x)) and P(y) is well-defined in terms of the pair (P, D), the relation "probabilistic cause" is, likewise, well-defined.' (Pearl 2001: 37)

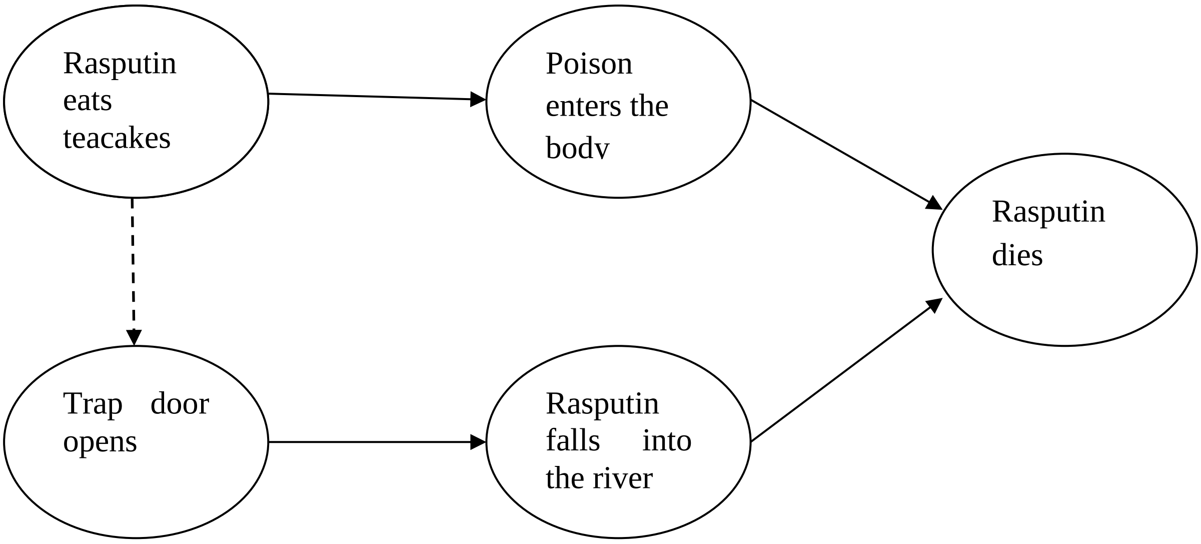
The main difference between Pearl's and Woodward's approach to causation that I am going to emphasize for my current purposes is that the former is a mathematical language built to deal with causal concepts, whereas the latter is meant to be a wider, philosophical account on causal relations, and extended to causal explanation.[26] Another important difference to stress, which will come about in the next section, is that Pearl is not talking about causal relations *strict sensu*, but about ways of inferring causal relations. If Woodward is giving a broader account of causation, then his notion of causation as manipulability needs a stronger backup than the claims found in Pearl.

Pearl's (2009) updated structural theory of causation, meant to unify the approaches to causation based on graphic, potential outcome, probabilistic, decision analytic, and structural equations models raises a few interesting issues that I am not going to discuss. Briefly, Pearl is showing how his approach based on the formal apparatus previously presented can include other approaches to causation. Woodward's approach itself puts together an interventionist conception, counterfactuals, and the apparatus of causal modelling. As I will consider the issue of the unification of the conceptual frameworks which have been proposed in defining causation in a more metaphysical way (as in Strevens 2013), I will not go into further detail about Pearl's unification approach.

A final thing to note in relation to Woodward's manipulability theory and the features of Pearl's work on causal inference is that the strength of Woodward's approach in dealing with various counterexamples relies a great deal on its structural features, even more so than on manipulability itself. This observation is made by Strevens (2008), where the following example is discussed: some conspirators invite Rasputin for tea and serve poisoned teacakes.

[26]Although, given Pearl's talk of deep understanding and intervention, it may be inferred that, as Woodward, he sees explanations as specifying what would happen given potential changes in causally relevant variables.

In case Rasputin decides not to eat the cakes, the conspirators build a trap that will open up if the cakes are left untouched for a certain amount of time. In this case, the trapdoor will open and drop Rasputin into the Neva, where he drowns. Thus, there are two causal chains leading to the same outcome, the death of Rasputin, illustrated as follows:



The issue at stake here is whether there is counterfactual dependence between Rasputin eating the cakes and his death. On a simple counterfactual account, if Rasputin had not eaten the cakes, he would still have died because of the events from the alternative causal chain occurring. On Woodward's account, on the other hand, intervening in order to prevent Rasputin's eating the cakes involves breaking all the arrows that depart from this variable. This means that, while Rasputin does not die from poison, the backup causal chain is also cancelled, since the connection between Rasputin not eating the cakes and the opening of the trap door is cut. There is a point to note here, however. As Strevens points out, 'the ability of the manipulation account to handle preemption has little or nothing to do with its manipulationist element: what works his magic is not the fact that the causal potency of the putative cause c is assessed by manipulating c's occurrence, but a logically separate element of the test, namely, its holding the off-path elements to their actual value.' (Strevens 2008: 63)

A conclusion to draw from this observation is that the strength in handling counterexamples should not be assimilated to the manipulability component only, and that the formal apparatus plays an important role.

1.1.4.4 Bayes nets and epistemic causation

Williamson's (2005) lengthy study on Bayes nets results into an approach that takes both probabilities and causal relations to be mental entities. His main objection against approaches that take both probabilities and causation to be physical is that they cannot deal with the wide array of counterexamples to the Causal Markov Condition. While Williamson discusses counterexamples showing that there are non-causal inducers of probabilistic dependencies, such as accidental correlations, relatedness through meaning, logical connections, mathematical connections, non-causal physical laws, I will not deal with them. Hausman and Woodward (1999) defend the Causal Markov Condition against several kinds of counterexamples and connect it to the concepts of manipulability and invariance. Woodward (2003) relies on these arguments in his discussion of causation and probability.

Williamson further shows that an objective Bayesian interpretation of probabilities would satisfy the Markov condition and thus allow for the use of Bayes nets. His definition of objective Bayesianism is 'that probabilities are an agent's rational degrees of belief (and so are mental entities) and these degrees of belief are fixed as a function of the agent's background knowledge (and so are - objective).' (Williamson 2005: 1). Taking after Bernoulli's work, he specifies that in objective Bayesianism both logical, as well as empirical constraints are taken into consideration.

The next step is opting for an epistemic view on causality. Williamson's definition for epistemic causality goes as following:

> The causal relation is mental rather than physical: a causal structure is part of an agent's representation of the world, just as a belief function is, and causal claims do

> not directly supervene on mind-independent features of the world. But causality is objective rather than subjective: some causal structures are more warranted than others on the basis of the agent's background knowledge, so if two people disagree about what causes what, one may be right and the other wrong. (Williamson 2005: 130)

Williamson is grounding this view of causation on two principles that, while not sufficient to back up a theory that assumes causal relations to be physical, nevertheless explain why the use of the concept of cause is useful. The principles are:

> **Qualified Causal Dependence** Normally causal relations are accompanied by probabilistic dependencies.
> **Strategy** Normally, instigating causes is a good way to achieve their effects. On the other hand instigating effects is not normally a good way to bring about their causes. (Williamson 2005: 137)

A further interesting claim that Williamson is making is that causal relations could be discovered through these two principles alone, and that, on his approach one need not show how common reasoning with causes has to rely on causal Bayes nets. His scheme for learning causal relationships includes hypothesizing, predicting, testing, and updating, using deductive, as well as inductive tools.

Having briefly described Williamson's view, there are a few things that could be pointed out connecting to some issues in Woodward's approach. First of all, Williamson's approach shows that the causal Bayes nets apparatus could be used to defend a theory that need not consider either probabilities, or causal relations as physical entities. Thus, there are several possibilities of connecting a theory about causation with the Bayes nets formalism, Woodward's manipulability view is one of them. As will become clearer in the next chapters, the advantages of defining intervention from a framework compatible with Bayes nets enable Woodward to deal with counterexamples better than other theories.

Second, since Williamson does not have a theory about the metaphysics of causal relations and, instead, he focuses on the instrumental value of these relations, the problem of inferring causal structures becomes important. In an approach such as Williamson's the question seems to be 'how do we infer causally?' rather than 'what do we mean by ‘causation’ when we make causal claims?’. Since Woodward is willing to offer a broader, philosophical perspective, he needs to go further into the metaphysics of causation. Relying on Pearl’s considerations is not sufficient since, as already pointed out, Pearl’s theory is about causal inference.[27]

Thirdly, a separate issue that both Williamson and Woodward raise is the common sense learning of causal relations. Williamson’s reliance on the qualified causal dependence and the strategy principle makes his approach compatible with a wide array of causal learning abilities. Woodward, on the other hand, as it seems to transpire from some of his articles about interventionist causation and investigations in the psychology of causal learning, seems to be committed to the claim that causal learning is based on the structure of causal Bayes nets, or at least that it fits his criteria for intervention.[28] I will explore this issue in detail in chapter 4.

Finally, intervention appears as one of the principles used by Williamson. He does not give an account of intervention as technical as Woodward does, so his principle can be interpreted as the common sense intuition that bringing about causes is a way of bringing about their effects, but not the other way around. He further specifies that this grants the asymmetry of causation. It is interesting to note that, at a first glance, Woodward seems to ground the asymmetry of causation in the asymmetry of intervention as well, and, furthermore, as shown earlier, a similar way of defining causation is found in Menzies’s and Price’s agency theory. The analogy goes as far as this, however as the metaphysics underlying these accounts are different: Williamson opts for an explicit epistemic view, Menzies and Price settle for an account of causation as a secondary quality, while Woodward endorses a realist approach. The further point that could be made from here is that intervention appears as an important component of a theory of causation, and also as an important way of making

[27] Although he does seem to hold the view that causal relations are physical.

[28] As I will be pointing out in chapter 4, this claim is endorsed in some of the psychology literature on causal learning.

sense of the asymmetry of causation. It will be an aim of the next chapter to spell out how effective the concept of intervention is in providing an account for the direction of causation.

To sum it up, the causal Bayes nets formalism provides a broader perspective over causation and probability, including the possibility of intervention along with observation. Among other things, this also makes it easier to deal with specific counterexamples. While Woodward's theory shares these aspects with the causal Bayes nets approaches, the choice between one metaphysical commitment or another is left open. As specified earlier in the chapter, Woodward opts for a realist view on causal relations; this, however, proves to have no bearing on the compatibility with the causal Bayes nets approaches, so it should be analyzed on philosophical grounds. My aim for the next section, as well as for the most part of chapter 2 is to investigate the metaphysical status of manipulability. Another point to make is that, in the absence of metaphysical commitments, the possibility of viewing manipulability as a causal inference, or learning theory is also available. I will discuss this in chapters 3 and 4.

1.2 Manipulability and metaphysics: the tension within Woodward's account

Up until this point, I have been looking into Woodward's definition of causation as manipulability and to related approaches. I have pointed out the way in which Woodward's account differs from the agency account through the commitment to realism. I have also shown that, although Woodward's version of manipulability shares the formal apparatus of causal Bayes nets, this does not in any way constrain his choice of metaphysics; such approaches are also compatible with a fully epistemic take on causation. I have pointed to the non-backtracking interpretation of counterfactuals as the most important similarity between

Woodward’s version of manipulability and Lewis’s counterfactual approach, while pointing to the metaphysics inherent to Lewis’s account of influence as compatible with the general picture resulting from Woodward’s definition of causation as manipulability. The question to ask here is what is the status of the manipulability theory of causation, metaphysically speaking? As the debate following Strevens’s (2007) review shows, things are far from straightforward.

Strevens’s (2007) review of *Making Things Happen* brings forth some criticism about Woodward’s definition of type-level causal claims, and about the claim that facts about causality metaphysically depend on facts about manipulability. By looking at what seem to be the metaphysical foundations of Woodward’s view, Strevens notes that all causal relations are built on the relation of direct causation. Woodward’s definition of direct cause holds that a variable X is the cause of another variable Y relative to a set of variables V if one can change Y through intervening X when the other variables in V are held fixed. This definition captures a situation where there is no intermediary between the two variables. However, this may become puzzling if one thinks of the world’s causal structure as being continuous. As Strevens raises the objection, ‘if the world's causal processes are continuous - in which case all causal relations are mediated - the notion of direct causation is only applicable relative to a V that represents a suitably sparse subset of the totality of causal facts.’ (Strevens 2007: 244) Under such assumption, it is hard to picture the notion of direct causation as fundamental.

Strevens also takes issue with the idea of the relativity of direct causation to a variable set. While, on a regular basis, causation is thought about without talking of variable sets, in Woodward’s theory there is no such non-relativized concept. I will discuss this objection in more detail in section 1.3.

Furthermore, Strevens does not share Woodward’s view on the supposed non-vacuously circular definition of interventions. In order to illustrate how vicious circularity comes in, Strevens proposes the following case: there is a variable set V containing X, Y and Z. In order to establish whether there is a causal connection between X and Y one must intervene on X through the intervention variable I. Given Woodward's definition of intervention, in this case one has to establish either that I does not cause Z, or that Z is not a direct cause of Y[29]. In order to establish whether there is a causal connection between I and Z one would need another variable. Even supposing that this would not lead to an infinite regress and such a variable were available, one would depart the initial variable set V, since the new variable set would have to include I along with X, Y and Z. If one is to see whether Z is a direct cause of Y, then one would need to see whether X is a direct cause of Y, which leads to a vicious circularity (the causal relation between X and Y is used in order to establish the very same causal relation). One way out suggested by Strevens is that Woodward should drop the interventionist view at the type level definition and thus 'to define intervention in terms of type level causation (…) but then to define type level causation in some way that makes no reference to intervention, or to take it as a primitive. This is, of course, to renounce a manipulationist metaphysics of type level causation.' (Strevens 2003: 245)

Further, Strevens holds that 'Woodward would be much better off (…) presenting his account not as a metaphysics or a definition of type level causation, but as describing something like a Putnam-style stereotype of our causal concepts, that is, as a theory of a kind of "content" that does not fix truth condition.' (Strevens 2003: 246) This would imply that one should pick another metaphysical view when defining type-level causation instead of using Woodward's concept of intervention: while the manipulationist view shows how we think about causation, it does not capture its nature. Even if one were to adopt this view, Strevens still has some objections. One of them concerns the relation between lower level and higher level causation. On Woodward's account, causation takes place between higher level relata.

[29] To recall Woodward's definition, I must not cause Y via a route independent from X (in this case, through Z).

This does not seem to imply anything about lower level facts constraining higher level instances of causation. For instance, one could talk about a certain state of the world causing another state of the world only under a specific configuration of the elementary particles. The manipulability account, as presented by Woodward, does not mention how such constraints may come into place. Finally, there are simple instances of causation, like tipping a glass and thus causing the water to spill that do not involve any sophisticated graph approach. Again, Woodward does not explain how such simple cases fit into his more elaborate picture.

Woodward (2008) rejects such criticism for most of its part[30] on the ground that *Making Things Happen* is not a book about the metaphysics of causation, but a contribution to the philosophy of science, namely, on how causal claims and experiments made by scientists are accommodated by the manipulability account. He also considers that, given his theory, he is not committed to any specific kind of metaphysics. To exemplify this:

> one of the attractions of the manipulationist account is precisely its unmetaphysical character—rather than thinking of causal relationships as involving mysterious other worldly entities (relations of necessitation among universals, similarity relations among possible worlds and so on), I urged instead that we think of them simply as relationships that are exploitable for purposes of manipulation and control. I argue that this makes it intelligible why we should care about discovering causal (as opposed to merely correlational relationships) and also helps to illuminate many of the ways in which we learn about and reason about causal relationships. For those who care about metaphysics, this sort of view might be supplemented by any one of a number of different stories about metaphysical foundations but MTH (*Making Things Happen* – my note) does not attempt to provide such foundations. (Woodward 2008: 194)

Another point where Woodward disagrees with Strevens is the assumption that an account of causation must have something to say about metaphysics. Woodward considers that, if one is to criticize an account of causation on the ground of its not dealing with metaphysical issues, one must have an argument for the claim that a theory of causation should incorporate some metaphysical foundations.

[30] The debate about the relativity to a set of variables goes on, I will come back to this issue in the next section.

In his reply, Strevens offers three arguments for giving Woodward's book a metaphysical reading:

a) While defending his theory, Woodward deals with explicitly metaphysical approaches such as the agency theory proposed by Menzies and Price, or Lewis's counterfactual theory. If Woodward presents his account as an alternative to these accounts, then he is implicitly claiming that his theory could fulfil the metaphysical tasks explicitly assumed by these theories.

b) Woodward's argument for the objectivity of causal claims (i.e. the employment of counterfactuals, through which intervention goes beyond an agent-relative perspective) makes sense only when interpreted as telling something about the nature of causation, and thus resorting to metaphysics.

c) Woodward states that his account is meant capture the meaning of causal claims and he employs definitions for this project. This commits him to the use of truth-makers for causal language, and, consequently, to a metaphysics of causal facts.

While I find Strevens's points convincing, there are a few things to mention here before moving on. First, that Strevens goes further into details about what Woodward's account is, if not a metaphysics of causation. The answer seems to be found Woodward (2014) where Woodward argues for manipulability as a functional account of causation, on which I will be focusing in chapter 3. I would like to start the next section by pointing to an additional reason why Woodward's concept of manipulability has an important connection to the metaphysical problem of causation and also emphasize the consequences this has on the debate over causal realism.

1.3 Manipulability and the realism-projectivism debate

In this subchapter I will come back to the realism – projectivism debate, which I only sketched earlier when discussing Woodward's view and the agency account. The main question I wish to address concerns what is at stake when one is presented with a choice between causal realism and causal projectivism. In what follows I will be discussing two arguments that favour realism, both used in some form by Woodward, along with a counterexample by Price that is meant to challenge the realist view. The first argument, through which this subchapter connects to the previous section as well as the inquiry into metaphysics from chapter 2, concerns explaining the success of intervention–based uses of causal reasoning, and as I will be arguing, it takes the form of the 'No Miracle Argument' from the literature on scientific realism. The second argument relates to realism providing grounds for objectivity via the correspondence with features of causal relations in the world, whereas on a projectivist view there is the question whether the dependence of causation on the agent's perspective would undermine the objectivity of causal claims. Finally, Price's counterexample relies on the idea that the agent's perspective would prove to be constitutive of causation in case it would be possible to switch between worlds where time would run in opposite directions. I will discuss these arguments in what follows.

1.3.1 Realism and metaphysics

In several articles (Woodward 2007, 2011a, 2012), Woodward argues for the manipulability approach to causation as an accurate account for the way in which causal relations are inferred in tasks concerning causal learning and causal reasoning in psychology. While I will look into this issue in more depth in chapter 4, right now I will focus on a question that Woodward answers when pointing to the compatibility between the manipulability account of causation and the psychological studies into causation. The question concerns what the connection between causation and causal inference is, or, to put it another way, what connects the philosophical inquiry into the concept of causation to experimental data? The answer is in line with Woodward's allegiance to causal realism:

> the two sets of issues—the worldly one about what causation is and the psychological ones—are closely interconnected. This is partly because (…) we would like to explain the patterns of success and failure in various tasks probing the nature and extent of various species' causal knowledge or competence. It is a natural assumption – made by many researchers and one that I shall accept – that successful performance means that the subject is in some way tracking or exhibiting a sensitivity to some features of causal relationships as they exist in the world (features that are relevant to success on the task) and that failure is an indication that those features are not being successfully tracked. (Woodward 2011: 4-5)

In Woodward 2012, describing his account as normative holds that 'if we have a normative theory that tells us that we ought to reason about causal relations in certain ways (...) and if we find people in fact reasoning in some good approximation to what is recommended, then these facts can form part of a potential explanation of why (and to what extent) people are successful causal reasoners. (Woodward 2012: 962-963)' In the light of the previous considerations, once again, the success at causal reasoning tasks can be attributed to reasoning in accord with the normative theory, which, in turn is grounded in some features of causal relations as they exist in the world.

This talk in terms of success or failure, and their explanation through the features of causal relations in the world is to a great extent similar to the talk about realism in the philosophy of science. Philosophers supporting a realist view make the claim that if scientists would not talk about entities out in the world, and would not pursue the truth, then one would not be able to explain the success or failure of various theories. Woodward seems to say the same about causal relations in the world and the psychological mechanisms of inferring causal relations.

To illustrate this, one could have a look at the ‘no miracle argument’ (see Putnam 1975). In a reconstruction by Chakravartty, the main point goes as follows:

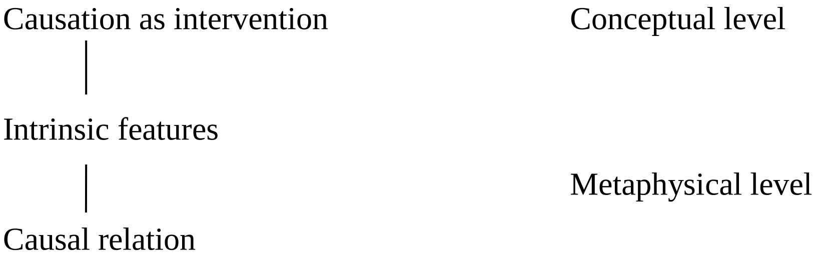
> The argument begins with the widely accepted premise that our best theories are extraordinarily successful: they facilitate empirical predictions, retrodictions, and explanations of the subject matters of scientific investigation, often marked by astounding accuracy and intricate causal manipulations of the relevant phenomena. What explains this success? One explanation, favoured by realists, is that our best theories are true (or approximately true, or correctly describe a mind-independent world of entities, properties, laws, structures, or what have you). (Chakravartty 2013)

Considering causal relations and leaving all the other issues regarding scientific realism aside, one can notice the same driving idea: it is the presence of causal relations in the world that makes our working with causal claims successful. The success of a certain kind of causal inference is explained by its use of structures that are also found within the nature of causal relations. I would like to clarify the fact that, although I am in favour of causal realism, it is not my purpose here to assess this particular way of argumentation; I am merely pointing out its similarity with the assumption that Woodward makes explicit. If one were to point out the weaknesses of this approach, or of this particular argument, I believe that the considerations of internal realism (Putnam) or quasi-realism (Blackburn) could be applicable, by extension, to any worries of the sort regarding realism about causation. One main point that I wish to make here, however, is that once this kind of argument is brought forward to connect philosophical and empirical issues about causation, one cannot remain silent about the metaphysics. The question to ask concerns what causal relations amount to. If manipulability is not a primitive feature of the world, and we have seen previously that Woodward claims that it is not, then what is it that renders causal judgments based on manipulability true? A further question concerns the nature of causation and how the manipulability theory captures its features. A look into metaphysical foundations may also be of help in finding a way of overcoming some shortcomings that manipulability does not deal with.

One worry that I would like to dispell at this point concerns Woodward's considerations on metaphysics and his functional project. I would like to focus on a claim that concerning realism: 'some may favor a very expansive conception of metaphysics according to which even this minimal realism (indeed any claim about what exists) amounts to a metaphysical commitment. But this unhelpfully makes every empirical claim a matter of "metaphysics."' (Woodward 2014: 698) While I acknowledge that empirical matters are to be discussed on separate grounds, I do not agree with Woodward on the complete separation of metaphysical and functional issues about causation. Particularly, I take such approach to fit in with a mainly epistemic or instrumentalist view on causation, where metaphysical foundations are not deemed to be important. A commitment to realism provides a satisfactory account of the connection between what causal relations are and how they are useful, which may lack from most epistemic/instrumentalist views, but this comes at the price of making claims about additional metaphysical issues. Rather than considering that such a view turns everything into metaphysics, I believe that looking into the metaphysical foundations of manipulability can answer interesting questions about the nature of causation and point out how ideas concerning manipulability can connect to a realist picture of causation. Furthermore, I would argue that Woodward's functional account is hardly metaphysics free. Two aspects of the manipulability theory whose role seems connected to metaphysics will be important for my discussion in the following chapters. One of them is the objectivist definition of intervention, which may be taken as the fundamental difference between Woodward's approach and other approaches to causation as intervention. Uses of interventions in order to single out causal relations have to meet all the demands that Woodward's definition of intervention specifies in order to qualify for the interventionist counterfactuals framework. Secondly, although he avoids the talk about truth-makers for such counterfactuals, Woodward opts for an explicitly non-backtracking interpretation of counterfactuals. From a metaphysical point of view, an important motivation is preserving the asymmetric nature of causation. From a functional view, as I will be pointing out in chapter 4, this take on counterfactuals excludes some everyday uses of counterfactuals in causal contexts.

Finally, the other important point that I want to make is that an argument along the lines of Woodward's argument presented above may give causal realism an advantage over projectivism as well. For this, I will have a look at the metaphysical and conceptual levels and their role in the agency and interventionist account, respectively. On both accounts, causation is defined in relation to some physical properties of the world, discussed in terms of intrinsic features. Manipulability, defined either in an agent-dependent or agent-independent manner, respectively seems to rest at the conceptual level. Thus, on both accounts the definition of causation relies on both objective features of the world and on a concept of manipulability. On the interventionist account, intervention is used for defining causation at a conceptual level, while the intrinsic features serve as objective grounds for connecting intervention to causal relations as they are in the world. On the agency account, however, agency seems to play an important role on both levels (being constitutive of causation), with intrinsic features also accounting for the objectivity of causation as agency.

Woodward's account of causation – intervention comes in at the conceptual level:



Causation as intervention

Intrinsic features

Causal relation

Conceptual level

Metaphysical level

Menzies and Price’s account causation – agency is constitutive of causation:

Causation as agency — Conceptual level

Intrinsic features + Human perspective — Metaphysical level

Causal relation

As the argument quoted above goes, the successful use of intervention in causal reasoning tasks can be explained through the fact that interventions exploit some features of causal relations as they are in the world. The projectivist view, although also relying on intrinsic features adds the human perspective to the concept of causation. While the success in causal tasks can be explained through the worldly features of causal relations there still seems to be a question about the relation between the human perspective and causal relations in the world. Given that agency is constitutive of causation, what guarantees its connection to the relevant features of causal relations such that people are typically successful in inferring causally? After all, people make all sorts of projections onto reality with more or less successful results. I will come back to this question after dealing with the objectivity issue in the next section.

Before moving on, I would like to come back to the earlier worry about projectivism, common causes, and ruling out confounders. As pointed out above, on the agency account the agent’s perspective is constitutive of causation along with the intrinsic features of causal relations and, as I will be arguing in the next section the agent’s perspective need not incorporate the experimenter’s beliefs or expectations, but some features that all agents share. One of the claims that I wish to make is that, even if the intrinsic features do not distinguish between causal and merely correlational relations, the agent’s perspective need not be eliminated from the picture when trying to avoid confounders. The reason for this is that the agent's perspective can be understood in a more objective manner, and thus, it does not necessarily incorporate thoughts or expectations by the experimenter. At this point one may bring in the desired methodology for assuring the reliability of an experiment. I will argue for the above mentioned objectivist take on the agent's perspective in the next section.

To sum it up, realism can provide a satisfactory account of why causal claims are useful. However that can only be done through looking at the metaphysical foundations for causal realism. As I will be pointing out in the next chapter for the manipulationist version of causal realism, the metaphysical foundations are insufficiently explored. For the moment, though, I will be exploring another problem which may provide another advantage to realism over projectivism.

1.3.2 Realism and objectivity

Regarding the latter argument for realism, concerning objectivity, what I will be mainly interested in whether taking causation to be agent dependent (either in its metaphysical or conceptual version) entails the fact that conflicting causal claims made by agents with different capacities should be taken as true. While there may be other reasons for why one would want causation to be defined in an objective manner, I take this to be one of the biggest issues with projectivism/the agency account. From this perspective, the distinction that will be relevant for my discussion is not about objective or subjective views on causation, but rather about how can a view taking causation to be agent dependent not collapse in the kind of subjectivism described above. In this section, I will be arguing that projectivism, and the agency account in particular, can provide a satisfactory answer to this objection.

One thing that I would like to draw attention to is the distinction between having a definition that allows for a certain concept to be objective and having a definition that involves a degree of dependence on a human agent. As specified earlier, Woodward's criticism of the agency account draws attention to the reliability of causal claims when they are in some way dependent on the capacities of the human subject. The point that I will be making, and which, as I will show, has its fundaments in Price’s (2007) perspectivalism is that objectivity and human-independence need not be mutually exclusive.

Price’s perspectivalism brings forth the idea that the epistemic position of the agent and the process of deliberation are constitutive for certain concepts such as causation (Price 2007), or temporal direction (Price & Weslake 2009). While there is no detailed account in his work on how perspectivalism fares with respect to objectivity[31], it is worth mentioning Price’s brief considerations on the concept of homogenous perspective. With respect to causation, ‘for basic physical reasons, all humans share a homogeneous perspective (...)’ (Price 2007: 251). But what should the homogenous perspective contain such that it would not lead to the concept of causation collapsing into complete subjectivism? I believe that Price’s considerations on the homogenous perspective can be used in an argument for the objectivity of a projectivist view on causation as follows:

1. Causation is dependent on the agent’s perspective. (perspectivalism/the agency view)
2. The agent’s perspective is constituted by those human capacities that are uniform across agents. (homogenous perspective)
3. Therefore, causation is dependent on those human capacities that are uniform across agents. (from 1 and 2)

Adding a few more assumptions about truth conditions, the conclusion is basically equivalent with the claim that the truth of causal claims is not dependent on goals, beliefs, or capacities that vary with different agents, which is loosely the concept of objectivity that I am arguing for.

The capacities enounced at (3) refer to the agents’ epistemic situation and their ability of making decisions. According to Price (2007), the agent’s perspective is characterized by

[31] I should mention, however, Price’s (2007, 2014) case against causal realism. Since in this section I will only be interested in one particular aspect of the disagreement between realism and projectivism, namely, objectivity, I will leave the discussion for the next subsection.

the deliberation situation where there are Fixtures (which can be either Known, or Knowable) and Options (which can be either Direct, or Indirect). Since every agent is aware of some fixed factors and some options for which the agent deliberates, this capacity is characteristic to all agents. In relation to causation, the core idea is that, even if causation is taken to be agent dependent, the way in which agents are described can be objective, singling out common traits of the agent. The objective description of the agent's perspective confers an acceptable degree of objectivity to the agency concept of causation.

Coming back to Woodward's quote above, a causal claim made as a result of a controlled experiment would not be undermined if the causal relation in question is partially dependent on the experimenter's capacities as long as those capacities are universal among agents. If, according to the Menzies-Price version of projectivism, causation always involves the agent's perspective, the reference to a specific human capacity is inevitable. Furthermore, the above mentioned approach to the agent's perspective can also be used to answer the methodological interpretation of this objection: even if the intrinsic features that ground causal relations are the same as those that ground mere correlations, the agent's perspective, which is meant to distinguish causal relations from correlations on the agency view, need not involve subjective thoughts by the experimenter.

Up until now I have been constructing a positive account of how a projectivist view on causation, namely the agency theory, can explain the objectivity of causal relations. For the remainder of this section I will go on to argue that the agent-independent definition of intervention in Woodward's own version might not meet the agent-independent concept of objectivity that it advocates.

As Strevens (2008) points out, Woodward's concepts of direct cause and contributing cause are defined by reference to a variable set. In Strevens's view, this leads to the undesirable consequence that 'what causes what depends on your perspective (more exactly, on the variable set singled out by your perspective).' (Strevens 2008: 174) While I will not go into the full details of the debate, I would like to emphasize a few more general points resulting from both Strevens's objection and from Woodward's response. Among other things, Strevens makes the point that the interventionist concept of direct causation represents only a part of causal reality. Since other variables could be added to the set and interfere with the direct causation relation, direct causation is taken to be relative to a variable set. Woodward's (2008b) answer is that 'perhaps the aspiration of the metaphysician of causality is to find a form of description that represents ''all'' of ''causal reality'' in a complete, non- partial way that is untainted by any purpose-relative human concerns (i.e., the sort of description that God would produce, if only He existed) but this isn't my project.' (Woodward 2008b: 211). Woodward also adds, contra Strevens, that this does not entail that the choice of variable sets is arbitrary.

There are a few points that I wish to make with respect to this debate. One of them is that, although not explicitly, on the passage quoted above, Woodward acknowledges the fact that the choice of variable sets depends on the purposes of one's inquiry. Subsequently, direct causation is not completely independent from the agent's choice. Furthermore, when rejecting the 'God's eye' viewpoint, Woodward seems to go one a step closer to Price's perspectivalism, which would not work either if the agents were all-knowing and could not engage into deliberation.[32] While it is essential for Woodward's account that the choice of variable sets not be arbitrary, a point which I will take for granted from his defence[33], I do not see how acknowledging the importance of human concerns when choosing variable sets can be given a completely human-independent interpretation. Does that threaten the objectivity of the interventionist account of causation, though? Before addressing this question, I will briefly reply one potential objection that a defender of the interventionist account might raise.

[32] See Price 2007: section 9.
[33] See Woodward 2008b: 206-211.

As Henschen (2015) points out, Woodward can reject Strevens's particular formulation of the objection (namely through the case of fine graining) on the grounds that fine-graining cases can be resolved by resorting to the notion of contribution cause. Nevertheless, Henschen further points out that the problem of variable choice is not limited to fine graining, and presents a case from the special sciences where 'researchers can differ with respect to what they are prepared to accept as serious possibility' (Henschen 2015: 7), ending up using different variable sets where different variables are connected through type-level direct causation. Thus, the variable set relativity of direct causation is a problem for the interventionist account as far as agent-independence as a mark of objectivity is concerned.

I was previously arguing that objectivity does not conflict with causal concepts being partly dependent on a human agent; therefore Woodward's concept of direct cause can retain an acceptable degree of objectivity such that the choice of variables, and thus, the claims about direct causation do not vary arbitrarily, but only follow different research purposes. While unlike the agency account, Woodward's concept of intervention manages to do away with the human agent, the human perspective comes back in the definitions of direct and contributing causation. It seems to be, thus, the case, that Woodward's own account does not meet its own standards for objectivity. However, if what I have been arguing previously is right, that should not threaten a more permissive concept of objectivity, including a human perspective.

An interesting issue to address here is whether the argument presented in the current section can be of use in defending projectivism against the claims made in section 1.3.1. If both realism and projectivism can account for an acceptable degree of objectivity for causal claims does that not justify the successful use of causal claims? My answer amounts to separating two questions, one about the degree of objectivity of causal claims under projectivism, the other about how would the successful use of causal claims be explained under projectivism. In this section I have argued that causal projectivism need not collapse into complete subjectivism. Nevertheless, even if causal claims under a projectivist view work, their success has to be explained, to some extent, by reference to the human perspective. As pointed out by Woodward in a passage quoted above, there is a normative component to successful causal reasoning. This component can be easily explained under the realist presupposition that causal claims based on interventions correspond to structures of the world. When structures of the world are mixed up with the agent's perspective (as objective as it could be) it is difficult to explain what that perspective corresponds to and why it works.[34]

Another interesting question to raise concerns the case against Woodward's relativization of the concepts of direct and contributing causation, namely, does that threaten his claim to realism? While this issue deserves more investigation, I would like to point out that, for the purposes of connecting manipulability to realism, having an agent-independent definition of intervention is very important. Furthermore, even if what proves to be a direct or contributing cause depends on one's choice of variables, the way in which the causal relations are supposed to work relies on some features of the world, exploited by manipulability. Thus, although distorted claims about causal reality can be made due to an arbitrary choice of variables, such claims do refer to causal relations 'out there' and their inadequacy is simply a matter of organizing the causal knowledge. Furthermore, as the recent literature on causation shows, the problem of finding derelativized concept of causation is open. For instance, Henschen (2015) proposes the follwing modifications to the definition of direct causation:

> X is a direct type-level cause of Y iff there is a possible intervention on X that will change Y or the probability distribution of Y when one holds fixed at some value all variables Z inside a set V of pre-selected variables and all variables R outside V. (Henschen 2015: 7-8)

[34] It would be possible to bring into discussion Williamson's talk of an ideal causal reasoner. The same question seems to apply to this approach as well: what is it that makes such agent ideal?

Without discusssing this approach to direct causation, the point I wish to make is that the project of connecting manipulability to causal realism could be pursued if various concepts of causation could be defined independently of the agent's choice.

Finally, it is quite likely that the concpet of objectivity that I have been arguing for may not satisfy the demands of 'hardcore' realism supporters. Then again, neither, would Woodward's concept of direct and contributing causation. In this respect, the issue of connecting manipulability to causal realism remains an open project.

I will now move on to an objection that may arise from the projectivist point of view, against realism.

1.3.3 An objectivist's dilemma?

Price (2007; 2014) opposes his perspectivalism about causation to the fully objectivist view attributed to Woodward. The dilemma he presents is supposed to have the objectivist choose between accepting that causation is ultimately tied to the agent's perspective or confront the threat of scepticism. While there are different specific scenarios to illustrate the counterexample, I will present the one from Price 2014.

Price (2014: 20) envisions a scenario where our world is linked to a distant spacetime region through two wormholes. The two regions have opposite temporal parity, which is to say, in the distant spacetime region times runs in the opposite direction from our own. In such a situation, causal claims and, implicitly, claims about what can be manipulated though what would depend on which wormhole one used. According to Price, this scenario shows that 'the extension of our anthropocentric notion of causation to regions in which we cannot actually manipulate things is in principle always provisional, and subject to correction in the light of learning more about the relevant physics.' (Price 2014: 21) The larger point that Price is making against Woodward and, by extension, against any view that aims at more objective perspective is that extending our concept of causation from our spatio-temporal region to a distant galaxy, for instance, is available only if one takes the perspective of the agent into account. In the case of this particular counterexample, what causes what depends on the agent's positioning next to one or the other wormhole. If one is unwilling to accept this anthropocentric feature of causation, then one might end up claiming that there is no causation in the distant region.

I would now like to address the question whether there is a satisfactory reply to this counterexample from the realist's side. Although a complete answer can be provided after addressing the issue of temporal direction and its relation to causation in the next chapter, for the moment I would like to point out that, in the particular example what makes our present concept of causation problematic for the distant spatio-temporal region is its relation to temporal order. Should this connect to the agent's perspective? According to Price, it should. Price and Weslake (2009) argue that temporal direction comes from the deliberative situation of the human agent, in pretty much the same way as causation, being a perspectival concept as well. Thus, on a perspectivalist view, causation in a region with opposite temporal parity to ours does not make sense if defined in a fully agent-independent manner. I agree with Price that Woodward's counterfactual definition of intervention cannot account for extending our concept of causation to this scenario (as in, one could only intervene to change the past, but not the future). I do believe, however, that an objectivist answer is available if a connection between causation and temporal direction is acknowledged. This solution is available for the realist as long as there is an objectivist approach to temporal direction. Such options, along with their compatibility to an objectivist definition of causation a la Woodward will be investigated in the next chapter. For now, my answer is that indeed our present concept of causation would not be operational in the said region. The extended concept of causation should incorporate the particularities of temporal order in the specific region and those particularities can be spelled out in an agent-independent manner. This solution, while keeping the objectivist aims, does acknowledge that causation is dependent on the nature of time.

One last issue that I would like to mention here concerns the status of the Menzies-Price account along with Price's updated considerations on perspectivalism and the Woodward account, respectively, as coherent theories of causation. While, as pointed out earlier, Woodward emphasizes realism and objectivity, his denial of metaphysics leaves some important questions unanswered. The agency account, however, presents a coherent view on causation, causal asymmetry, temporal order, and the human perspective. In the next chapter I will explore some options for a realist version of the manipulability account to account for these issues.

Chapter 2: Manipulability and the asymmetry of causation

In this chapter I will be looking at ways of accounting for the asymmetry of causation from the perspective of a manipulationist theory. Starting the inquiry with Woodward's approach may be instructive for a wider project of investigating how a manipulability theory of causation tied to causal realism might deal with the asymmetry issue, and what problems it needs to face. As with the previous chapter, I will be contrasting the options of the realist version of manipulability to the projectivist one and assess the available solutions. Since this is by and large a metaphysical issue, the inquiry into Woodward's account might need some clarification. As specified in the previous chapter, although Woodward denies any metaphysical aims for his approach to causation as manipulability, there are reasons to look into the metaphysical foundations for manipulability, and furthermore, to explain how some features of Woodward's version of manipulability relate to particular stances concerning metaphysics.

While I will be addressing an issue that generally concerns causation as a metaphysical problem, the causal asymmetry could also be investigated on the grounds of conceptual analysis. The metaphysical aspect of it is that causal relations follow a direction (from causes to effects, but not the other way around) and there should be a basis for that. The conceptual take on this issue would be that, when used in causal claims, causes are taken to make a difference to the occurrence of their effects, whereas effects do not make a difference to the occurrence of their causes. Thus, even if one might not care about the metaphysical foundations, an explanation of why causal claims are used in such a way in epistemic contexts is necessary. Both issues will be important for the points that I will be making, although this chapter will focus on the metaphysical perspective for most of its part. To sum it up, the main question that I will be addressing in this chapter concerns the source of the causal asymmetry.

My answer will point to the direction of time in both a stronger (metaphysical) as well as weaker (concerning causal understanding) claim. The way in which these two claims are connected goes back to the argument presented in the previous chapter, regarding the connection between what causation is and how people get to know causal relations. A realist take on this issue would result in connecting these two aspects.

I will start by showing that Woodward's definition of causation as manipulability does not provide a satisfactory answer to the question why causal relations are/are thought of as asymmetric. I will then look at the solutions provided in the literature on the direction of time and causation and assess their compatibility with manipulability.[35] Finally, I will present the solution that opt for: using the direction of time for explaining the direction of causation. I will further discuss the issues that this solution may bring about and how to answer them. I will also explain the two senses in which this solution could be applicable, as a metaphysical claim, and/or as a claim about causal understanding, with the aim of expanding the latter to Woodward's considerations on manipulability as a functional account of causation. The considerations that I will be making will also be of help in expanding on the argument presented in section 1.3.3 that could help realism reply to the projectivist challenge.

2.1 What explains the asymmetry of causal relations under the manipulability account?

While it may be fairly obvious that causal relations have to go from cause to effect in order for interventions on a cause variable to change an effect variable, the question is whether this asymmetry is the result of defining causation through intervention, or whether the asymmetry of intervention follows precisely from the asymmetric feature of causal

[35] That need not be limited to Woodward's version; there might also be other objectivist views on manipulability.

relations. I will go on and argue that on Woodward's account the latter claim holds. Furthermore, there is no explicit claim as to what may explain the causal asymmetry within Woodward's account. In order to clarify this issue there are two main places to look at, pertaining to the main components of Woodward's theory. One of them is the interventionist component; the other is the counterfactual one. While starting with the former, I will also explain that the objection that applies to Woodward's account is a problem for any account that seeks to ground the asymmetry of causation in the asymmetry of intervention.

2.1.1 The asymmetry of causation and the asymmetry of intervention. Which explains which?

One question that needs answering when looking at the concept of intervention and the asymmetry of causation concerns the metaphysical foundations. As discussed in the previous chapter, Woodward seems to be aiming for an account of the concept of causality. However, given the commitment to realism, Woodward's manipulability account could also be interpreted as more than an account of the concept of causation, that is, as incorporating some metaphysical claims. On either interpretation, the allegiance to realism brings about some questions about metaphysical foundations, one of which is the asymmetry problem. In this chapter I will be looking at the asymmetry problem from the perspective of both metaphysics and of causal understanding. The reason for doing so is that I believe there are interesting interactions between the two areas of inquiry which might be of help in clarifying the problem. However, I should note that were one to subscribe to a purely epistemic take on causation, then one need not look into metaphysics and the question regarding why people understand causal relations as asymmetric may be a purely empirical one.

Before looking at the particularities of Woodward's account it is worth investigating whether an account of causation based on a manipulability feature can account for the direction of causation. Such investigation has been done by Mackie, and the possibility of grounding the direction of causation (or, in his terms, causal priority) has been dismissed on grounds of circularity:

> If there is a causal relation between A and B, and we can control A without making use of B to do so, and the relation between A and B still holds, then we decide that B is not causally prior to A and, in general, that A is causally prior to B. But this means only that if one case of causal priority is known, we can use it to determine others: our rejection of the possibility that B is causally prior to A rests on our knowledge that our action is causally prior to A, and the question how we know the latter, and even the question of what causal priority is, have still to be answered. (Mackie 1965: 262)

And further, ‘it is true that our knowledge of the direction of causation in ordinary cases is thus based on what we find to be controllable, and on what we either find to be random or find that we can randomize; but this cannot without circularity be taken as providing a full account either of what we mean by causal priority or of how we know about it.’ (Mackie 1965: 262-263) The last point is particularly important since it emphasizes the fact that while people ordinarily associate the direction of causation with changing effects through changing their causes and not the other way around, this cannot explain people’s knowledge of this feature of causal relations. Thus, the argument also works against an epistemic perspective on causation as manipulability, and might be extended to counter the claims that the asymmetric feature of causal relations is learned through intervention: according to Mackie’s claim, while the direction of causal relations may be associated with interventions, its knowledge is prior to the knowledge about the results of interventions.[36]

Turning to Woodward’s account now, things seem to be particularly problematic since intervention is explicitly defined as a causal concept. As Mackie’s argument shows, not even a non-causal definition of intervention could offer a satisfactory account of the direction of causation. One question to ask here is whether one way out of this problem would be to admit circularity, but argue that it is informative, in the way Woodward does with his definition of

[36] It would be interesting to investigate this claim in empirical context, as in whether causal relations are understood as asymmetric and if so whether knowledge about the asymmetry comes before the use of information about the results of interventions. I will say more about that in chapters 3 and 4.

causation. Woodward's account, however, does not seem to explain why causal relations are asymmetric. His theory posits some intrinsic features of causal relations that make them exploitable for the purposes of manipulation and control. Maybe that would be the place where the asymmetry comes in, but Woodward's approach to manipulability does not describe such features. At the same time, it is clear from Woodward's definition of intervention that causal relations involve an asymmetry: effects can be controlled through their causes; causes cannot be controlled through their effects. The issue that I want to point out is that the asymmetry of intervention is explained through the asymmetry of causal relations, and not the other way around. Thus, it seems that Woodward's account makes use of causal relations already assumed to be asymmetric and does not provide grounds for the asymmetry of causation. While this problem may be dismissed along with the other metaphysical worries surrounding Woodward's approach to causation as manipulability, things do not fare much better from a strictly conceptual perspective. Our everyday concept of causation contains the idea that causation follows a direction, and the link between causation and intervention seems to be relying on it. If one wishes to find an explanation as to why the concept of causation involves an asymmetry between causes and effects, the explanation would need to make use of a concept in terms of which both the direction of causation and that of intervention can be explained. Since interventions are defined in a counterfactual manner, another place to look for an explanation would be Woodward's take on counterfactuals. While I will be doing that shortly, I would also like to quickly discuss another two options of connecting manipulability with the direction of causation.

One of them concerns the apparatus that backs up Woodward's concept of intervention. Since the arrow-breaking feature of manipulability renders Woodward's account able to deal with counterexamples, a further option to investigate is whether the formal apparatus defining intervention might be of help in explaining why on the manipulability view causal relations are asymmetric. However, as mentioned in the previous chapter, the compatibility with the causal Bayes nets framework does not lead to any metaphysical commitments from the perspective of Woodward's version of manipulability. Subsequently, it cannot explain why causal relations are asymmetric. The arrows in a causal graph are indeed directed from causes to effects, but that is a feature of the system, which does not entail anything about the worldly characteristics of causation. What about the way in which people use causal concepts and infer causal relations? Does the causal graph approach impose an asymmetric view on causal relations? While I will pursue this issue in more detail chapter 4, for now, I am only mentioning that it is unlikely that people make use of such a complex apparatus every time they reason about causes, let alone of the particular manipulationist version that Woodward proposes. If there are to be some grounds for the direction of causation, they should be searched elsewhere.

The other possibility to look into is to take the asymmetric feature of causal relations as a primitive.[37] There are two main issues with this option. One is that it would provide little insight into how causation as an asymmetric relation and causal understanding work, at least from the more metaphysically-oriented approach that I endorse. Secondly, as I will be explaining later on, it is not at all clear whether taking causation as a primitive is compatible with a manipulability definition that ends up relying on the human perspective at one point or another.

Finally, before moving on, I should also emphasize that Mackie's argument is a threat for other attempts to explain the causal asymmetry through a manipulability feature as well. Since I have also been getting into the projectivist version of manipulability, a question to answer is how such a theory can avoid this objection. I will postpone this discussion for section 2.2. I will now look into the counterfactual component of Woodward's theory.

[37] Frisch (2012) seems to suggest this with respect to Pearl's and Woodward's concept of intervention. I will argue that this might be a problematic claim in relation to Woodward's account. Also note that this option would be compatible with the causal graphs approach.

2.1.2 Backtracking counterfactuals and the asymmetry of causation

The problem of backtracking is one of the major issues that counterfactual accounts of causation have to deal with in connection with the causal asymmetry. While Woodward's account is not a purely counterfactual account, his use of counterfactuals when defining intervention may enable an explanation of the asymmetry of causation in terms of the asymmetry of counterfactual dependence. Such an attempt was previously made by Lewis and in this section I will mostly rely on the Lewisian approach to explain the issues around backtracking. In the previous chapter I pointed out that the non-backtracking interpretation of counterfactuals is among the most important similarities between Lewis's and Woodward's accounts of causation. The reason why I am illustrating the problem by reference to Lewis's rather than Woodward's work is that Woodward does not explain the truth conditions for interventionist counterfactuals (see Woodward 2014: 698-699). While this is associated with Woodward's non-metaphysical perspective on manipulability and causation, it does seem to leave the space open for criticism concerning backtracking counterfactuals. For someone willing to go further into the metaphysical foundations for manipulability, the issue is that whatever causal relations in the world are taken to be, they must support a non-backtracking interpretation of counterfactuals. In this section, I will look at Lewis's attempt to deal with this problem in order to emphasize the problem, as well as explain the further attempts at explaining causal and temporal asymmetries.

To recall, Lewis (1973) attempts to explain the asymmetry of causation as well as the direction of time through the relation of counterfactual dependence and the similarity between possible worlds criteria.[38] There are several problems with his treatment of these issues, most notably the problem of backtracking counterfactuals. I will mostly focus on Bennett's (2003)

[38] I have enumerated the criteria in Chapter 1.

considerations on the issue. According to Bennett, Lewis's main problem is to establish which counterfactuals are true while maintaining the causal and the temporal asymmetry. In Bennett's reconstruction, Lewis's account unifies forward and backward counterfactuals by relating them to the same set of worlds:

'A>C iff C obtains at the A-worlds that most resemble α at T_A, out of all the A-worlds that become unlike α for the first time at a late modest fork and are legal from then onward.' (Bennett 2003: 276)[39]

As Bennett further notes, 'in Lewis's theory many forward counterfactuals are true while relatively few backward ones are—merely ones reaching back down the ramp towards the fork—and he (…) offers this as explaining the meaning and securing the truth of the common idea that the future is open, the past closed.' Leaving the time issue aside for a moment, one could see at this point how Lewis's approach to backtracking counterfactuals relates to causation and counterfactuals. His theory, relying on the latest fork (i.e. the latest moment when a possible world diverges from the actual world) does not allow for the counterfactuals that go as far back as before the fork to be true.

His aim of grounding temporal direction in counterfactual dependence affects his analysis, and, as Bennett points out, makes it vulnerable to criticism. In order to have a non-circular account of temporal direction, Lewis must analyze the subjunctive conditionals in terms that do not involve temporal order. He does so by replacing what Bennett describes as "latest fork" with "as long as possible": thus, for instance, w1 is closer to the actual world than w2 if it is similar to the actual world for a longer time before a divergence occurs. Bennett further explains the motivation of Lewis's considerations on small miracles:

> Lewis absolutely needed the thesis that (roughly speaking) *small miracles cannot make worlds converge to perfect likeness*. Without it, he could not exclude from closest A-worlds miracles leading to wrong truth

[39] In Bennett's terminology, A stands for the antecedent, C for the consequent, > for the subjunctive conditional, and α for the actual world. He defines fork as 'an event in a A-world virtue of which that world for the first time becomes less than perfectly like α', and ramp as 'the segment of that world's history starting at a fork and ending at the obtaining of A' (Bennett 2003: 217)

> values for many conditionals, except by bringing in the concept of temporal order in the stipulation that at a closest A-world no miracle occurs *after* the time of the fork. (Bennett 2003: 293)

Bennett's counterexample is that there could be a world, w1, similar with α until T, when a small miracle renders w1 to converge with another world, w2. Although Lewis claims that in order to make a world re-converge one needs a big, widespread miracle, in this case a small miracle makes w1 and w2 converge. Bennett's solution is 'to say that a world close to α must sufficiently resemble α until shortly *before* T_A, have no large miracles, and have no small ones at times other than that of the first divergence from being sufficiently like α.' (Bennett 2003: 298) Of course, this would go against Lewis's aim of spelling out these conditions without using any terms in relation to temporal order.

This convergence issue is well illustrated in an example by Elga (2000). Elga proposes to think of a world α* having the same dynamical laws as α, while being the time reverse of α. At t*, closely before t, a small miracle occurs such that the world is led to the fork where Nixon pushes the button.[40] In this example, α* starts with very high entropy that decreases until time t where its configuration overlaps with α. Due to a small miracle, from that point on the entropy starts to increase, just like in the actual world. The problem with Lewis's similarity criteria is that such a word is as close to the actual world as the world where Nixon pushes the button. But since α* is a time reverse of α, it turns out that Lewis's considerations on possible world similarity fail to capture the asymmetry of time.

The above mentioned criticism shows that Lewis's analysis of the direction of time in terms of his asymmetrical analysis of forward and backward subjunctive conditionals cannot be maintained in its current form. An important point for the current discussion is that Lewis's similarity criteria are meant to answer the problem of the direction of time from the framework of Lewis's analysis of counterfactuals. This explains why the criteria might seem

[40] Assuming, like in Lewis's original example that w1, where a small miracle occurs at t, Nixon pushes the button and a nuclear holocaust occurs, is closer to the actual world.

ad hoc to someone pursuing the issue of causation: they are meant to rule out backtracking cases. Thus, as the criticism by Bennett and Elga shows, the asymmetry of Lewis's analysis of counterfactuals is not the same as the asymmetry of time, and it cannot be used to explain the asymmetry of causation either.

Coming back to Woodward's approach now, it is worth pointing out that Woodward, as well, opposes Lewis's similarity criteria, among others, on the ground that they seem ad hoc. Woodward's 'upgrade' of the simple counterfactual analysis to an analysis using interventionist counterfactuals connects to manipulability and control. As explained above, from a metaphysical perspective, intervention cannot provide a satisfactory treatment of the asymmetry issue. Furthermore, the more functionally oriented issue of how people learn about the asymmetry of causal relations is also unaccounted for. In the absence of an account for the truth conditions for counterfactuals, the backtracking problem still persists.

It might be argued along the lines of Woodward's denial of getting involved in metaphysics, that the direction of causation is part of the features of the causal relations in the world that manipulability rests on. While, as pointed out previously, Woodward does not pursue such a project, I believe the endeavour is worthwhile for someone who is seeks to connect causal realism and manipulability. Subsequently, assuming that some account of metaphysical foundations for manipulability is necessary, I will look into potential ways of explaining the asymmetry of causation.

2.2 Perspectives on the asymmetry of causation

In this section I will be looking at various ways of dealing with the causal asymmetry and investigate their compatibility with a manipulability concept of causation. The views that I will be discussing will be the third arrow approach by Loewer and Albert, the subjective view by Price and Weslake, Frisch's explanation of the direction of time through the direction of causation and Hausman's treatment of causal asymmetries.

2.2.1 The Lewis-Albert-Loewer objectivist view

One way of explaining the asymmetry of causation is to connect counterfactuals to a different way of making sense of the temporal asymmetry. Loewer's (2007) understanding of counterfactuals involves connecting the temporal asymmetry of counterfactuals to the second law of thermodynamics. This project is meant to offer a scientific account of the asymmetry that both Lewis and Bennett consider when dealing with counterfactuals.

Without going into the details of the scientific background, I will mention that Loewer's account of counterfactuals is in line with Albert's (2000) proposal of adding two claims to the fundamental dynamical laws. The claims are:

> (PH) a statement specifying the macro state of the universe at one boundary (which is assumed to be on with very low-entropy condition satisfying certain further symmetry conditions).
>
> (PROB) a uniform probability distribution over the physically possible initial conditions compatible with PH; i.e. the initial macro state of the universe. (Loewer 2007: 411)

Going on to a brief description of Loewer's analysis of conditionals now, the SM (statistical mechanics) account of counterfactuals involves decision conditionals of the form: 'If P were to decide to A then the probability of B would be x.' (Loewer 2007: 429). Decisions are characterized by two assumptions, first, 'they are localized events (or states) in a person's brain that are smaller than macroscopic events but have positive probability' and they 'are correlated with motions of her [the agent's] body' (Loewer 2007: 429). Thus, they end up being 'indeterministic relative to the macro state of the brain and environment prior to and at the moment of making the decision. This *indeterminacy* captures the idea that which decision one makes is "open" prior to making the decision.' (Loewer 2007: 429-430)

Decision counterfactuals come out to be true or false based on the statistical mechanical probabilities. To use Loewer's example 'the conditional "if I decide to bet on the coin's landing heads then the chance I would win is 0.5" is true at *t* iff the statistical mechanical probability of winning given the *t* macro state and my decision is 0.5.' (Loewer 2007: 430) These counterfactuals are asymmetric in virtue of the temporal asymmetry of the statistical mechanics distribution. While one's decision at t can make a difference to how the events after t occur, one cannot change the probabilities of macro events prior to t. In order to extend this account to non-decision conditionals, Loewer has to make use of the notions of past and future. He takes this to be unproblematic since these conditionals are based on SM conditionals: 'the proposal for evaluating non-decision conditionals is parasitic on decision conditionals.' (Loewer 2007: 434)

While there are several advantages of the proposal discussed above I will only mention one of them: the fact that it grounds the asymmetry of counterfactual dependence and time onto the laws of physics. Thus, when confronted with a question concerning why one should understand counterfactuals in this way, Loewer's answer is that 'the information expressed by SM-counterfactuals is important for us because it tracks the statistical mechanical probability distribution in ways that are important for the consequences of our decisions. (Loewer 2007: 438-439)'

There are some further issues concerning this proposal. For the present discussion, the most important is that a further project would be to connect the SM account of counterfactuals to a theory of causation. Given the fact that the SM account does not involve causal concepts and that the truth value of counterfactuals is supposed to match Lewis's intended outcome, such a theory of causation could be very close to Lewis's counterfactual account.

One question here would be whether this take on counterfactuals could be put together with Woodward's considerations on manipulability. One thing to consider would be that the

SM account of counterfactuals would provide an explanation why backtracking counterfactuals are false. Another interesting thing to emphazise is the reliance on decision counterfactuals. As making decisions is ascribed to agents, counterfactuals are understood in relation to agency. In an article about laws and time (Loewer 2012) where he calls the association between the fundamental dynamical laws, PH, and PROB, the Mentaculus, Loewer connects this solution to the asymmetry of control:

> On the Lewis-Albert the temporal asymmetry of explanation, causation, and counterfactuals is grounded in the probabilistic correlations of the Mentaculus. Why do we explain the future by the past and not vice versa? The Mentaculus answers this by connecting explanation and causation to control. The idea that causation is closely connected to intervention and control is widely held. (Loewer 2012: 132)

Thus, according to the passage quoted above, the Mentaculus solution could explain both the fact that causal relations are asymmetric and their connection with the asymmetry of one's capacity of changing the future, but not the past. There seem to be several issues worth discussing here in relation to the above mentioned problems. One question is how much does this specific sense of decision making or agency connect to Woodward's version of manipulability and its connection to causal realism? After all, causation is connected to intervention on the agency account as well. As specified earlier, one way of drawing a distinction between the two accounts is though Woodward's claim of objectivity for his concept of intervention. Given the background provided by Loewer's solution, the objectivity of causal claims no longer seems to be a problem. Furthermore, counterfactuals are understood by relation to decision-making. A question here would be whether, due to the use of counterfactuals and objectivity claims, Woodward's take on intervention is best suited for making sense of SM-counterfactuals, or whether different takes on manipulation and control could work as well.

Another issue concerns what causation is taken to be at the fundamental level. It appears that while the solution above has a similar take on the truth value of counterfactuals as the one by Lewis, at the fundamental level, causation seems to be defined by probability relations. Thus, this approach could connect Woodward's claims on manipulability to objective probabilities, since the Mentaculus relies on the objective probability distribution given by statistical mechanics.

Finally, on Loewer's account, the laws of physics play a crucial role in defining the asymmetry. If one were to supplement Woodward's difference-making criterion with such an account and expand it to scientific explanation, then one would have to drop the claim that the manipulability theory does not make use of laws in order to provide a theory of explanation.

At this point, a legitimate question would be why try to put together the two accounts in the first place? One obvious answer would emphasize the importance of Lewis's views on counterfactuals for the framework sketched above. I have already pointed out the similarities between Lewis's and Woodward's views. Another answer is that explaining the asymmetry of time and causation through the second law of thermodynamics and decision counterfactuals yields into similar truth-values for counterfactuals for both views. By connecting these two views, we are presented with an account of counterfactuals that fills in the gap left in the interventionist framework. Thirdly, this solution endorses an objective view that matches Woodward's account of intervention – although thought of in terms of decision counterfactuals, the asymmetry is given by the statistical-mechanical probability distribution.

2.2.2 The Price-Weslake subjectivist view

I am now going to point to some issues that the Albert-Loewer approach may have to face, and present the subjective alternative. In an article discussing the time asymmetry of causation, Price and Weslake (2009) present several solutions and their drawbacks. The objection they raise against explaining the asymmetry of time and of causation through the second law of thermodynamics and the past hypothesis involves considering a future hypothesis (FH). The question to ask here is if one were to know the end state of the universe, would that imply that one cannot affect the future? Price and Weslake point out that such hypothesis has almost no bearing on one's deliberation about one's future. Thus, if a future hypothesis would not completely prevent one from making changes to the future, a past hypothesis would not be able to do that either. The authors also discuss the possibility of having a closer future constraint. Using an example by Gibbard and Harper (1978), if one is destined to meet Death at noon on a certain day, even though most possibilities are closed, one can still choose the city where one will be at noon which, in turn, would affect Death's movements. As the authors point out, 'the example suggests that while a lawlike future constraint can limit the options, it does not make it absurd to think that we exercise control within those limits.' (Price & Weslake 2009: 426). Therefore, taking another look at PH, it is not at all obvious why it would explain why decisions are made towards affecting the future and not the past, or why effects can be changed through their causes, but not the other way around.

The solution that Price and Weslake advocate is to drop the 'third arrow', objectivist picture and think of deliberation in epistemic terms. As they put it, 'we have discovered that if we think of deliberation, initially, in epistemic, evidential or 'pre-causal' terms, it nevertheless exhibits a strong temporal asymmetry: an asymmetry explicable, apparently, in terms of our own asymmetric temporal orientation, as 'players' in the dynamical environments in which we live (...)' (Price & Weslake 2009: 436). The project they endorse implies that 'for causation (...) the practical, epistemic perspective is importantly prior to the metaphysical perspective.' (Price & Weslake 2009: 437).

An interesting, albeit orthogonal, issue to discuss here is Price's (1992) critique of the objectivist treatment of the asymmetry of causation. He mainly criticises Lewis's view on the grounds that in microphysics it is just as easy to converge to or diverge from a possible world.

Since I have been discussing several critiques of Lewis's account, the positive part of Price's article will be more interesting for the current discussion. His claim that 'the asymmetry of causation simply reflects (or better, perhaps, projects) that of the means-end relation' (Price 1992: 515). Thus, in conformity with his more recent work, the causal asymmetry is dependent on the agent's perspective. One problem that I would like to address here, mentioned in Price 1992, but dismissed, is whether the perspectival account for the asymmetry of causation is circular (as in Mackie's argument). In the light of the considerations from Price & Weslake 2009, as well as through its projectivist view on causal relations, I believe that the agency account of causation can avoid the objection. The issue of circularity is dealt with by taking decision-making epistemically and not as a causal notion: 'deliberation can be characterised in a non-causal, epistemic fashion.' (Price & Weslake 2009: 439) If people's picture of the world and causal relations is partly dependent on their experience as agents, then it is possible for them to write agent-specific features into the nature of causation or temporal relations. On this picture, causal relations are asymmetric, but they are no longer agent independent.

Coming back to the objectivist view, I will not go any further into the question whether, or how, the FH objection by Price and Weslake can be answered. The important point to make here is that although there are options for accounting for the asymmetry of time and causation, there are further problems that need to be solved. The question that I will try to answer here is whether the Albert-Lower or the Price-Weslake account of the time asymmetry of causation would be preferable as grounding the asymmetry expressed in Woodward's analysis of causation. Interestingly, both Price and Weslake, as well as Loewer (2007) refer to interventionism as support for their solution. The crucial question to ask here is whether Price and Weslake's characterization of intervention satisfies Woodward's 'modest realist' views.

I believe that the part of answer is already lies in the discussion on manipulability and agency in chapter 1 and in the previous sections of this chapter. First of all, Woodward distances his view from the agency account on grounds of causation being independent of the human ability of agency. His emphasis on counterfactuals is meant to replace the agent perspective present in the Menzies and Price account. Furthermore, his argument for connecting research concerning causation in philosophy and psychology relies on a realist argument: successful causal reasoning tracks features of the world. Rather than connecting the Price-Weslake account of the time-asymmetry of causation to Woodward's concept of intervention, I would suggest that it supports something in the fashion of the Menzies and Price agency account, or Price's perspectivalism: causation is connected with the human capacity of decision-making whose asymmetric feature is understood epistemically.

As for the positive arguments that favour connecting Woodward's concept of intervention to a third arrow view, I have already pointed out how objective probabilities and the SM account of counterfactuals could provide the metaphysical backup for the semantics of counterfactuals at use within Woodward's account. A final thing to note is, once again, the tension that I was describing chapter 1: the connection between causation and intervention can be used for both objectivist and subjectivist accounts of asymmetry as long as they involve decision-making. By opting for a realist view, Woodward's account seems to be closer to the objectivist treatment of asymmetry and counterfactual thinking.

2.2.3 Frisch on the connection between the causal and temporal asymmetries

I will now discuss an approach to the causal asymmetry and temporal direction by Frisch (2014). While Frisch's discussion mainly focuses on causation in fundamental physics, there are a few relevant points to mention for my discussion here. A first thing to note is that Frisch agrees with Woodward's functional view on causation[41] and he argues for the

[41] I will have more to say on that in the next chapter.

usefulness of causal claims in the domain of physics. He also supports a realist view where 'causal representations, like scientific representations in general, always are representations by us for a specific purpose in a certain context, without implying that how we successfully represent is entirely up to us.' (Frisch 2014: 11). While relying on causal graph representations and counterfactuals, in the manner of Pearl and Woodward, Frisch disagrees with the Lewisian (and, by extension, to what is implied by interventionist counterfactuals) interpretation of counterfactuals as time-asymmetric:

> [...] instead of attempting at grounding the causal or intervention asymmetry in a counterfactual asymmetry, I am proposing that we turn Lewis's account on its head and ground the asymmetry of intervention counterfactuals in the asymmetry of the causal model within which they are evaluated. Moreover, contrary to Lewis's view, I do not believe that there is a standard kind of context for evaluating counterfactuals in which backtracking counterfactuals are false. (Frisch 2014: 96)

These considerations help Frisch construct an argument for the use of causal asymmetries in order to explain the asymmetry of interventions on closed physical systems (see Frisch 2014: 101). Another point to stress in relation to the previous discussion of the objective and subjective treatment of the causal asymmetry is that Frisch opposes both the Albert-Loewer account of grounding the asymmetry of causation in counterfactuals based on the entropy, as well as Price's perspectivalism. While I will not go further into this debate, it is worth discussing Frisch's solution. Frisch's account does not aim at reducing causation to other worldly features coming down to thermodynamics. In his view 'both the causal asymmetry and an initial randomness assumption (...) are two aspects of a fundamental temporal asymmetry in the world that is reflected in our explanatory practices and in the representational resources we use.' (Frisch 2014: 201)

While Frisch agrees with some of Woodward's claims, notably the use of causal models, the realist assumption, and the functional project, an important question that arises is whether having a definition of causation as manipulability is compatible with taking causation to be one of the features of the world. As I pointed out in the previous chapter, when using a definition of causation as manipulability, the human perspective comes in at one place or another. If causation consists in some worldly features plus a human perspective, then it cannot be among the fundamental features of the world. This seems to be roughly the point made by Woodward (2007a). At this point the distinction between metaphysical and conceptual claims about causation could be handy. Arguably, neither Frisch nor Woodward are making strong metaphysical claims. Nevertheless, while Frisch argues for the use of causal terms in fundamental physics and to the causal asymmetry being constitutive of the temporal asymmetry, his claims mostly refer to the role of causation in scientific representation (particularly, in physics). The question here is whether the human face of causation would be problematic for causal representations in fundamental physics or only for the claim that causation is among the fundamental features of the world. Note that the latter, metaphysical question, is of great importance when trying to explain the causal asymmetry (i.e. is the causal asymmetry more fundamental than the temporal asymmetry, or is it the other way around?) While Frisch and Woodward may be running only conceptual claims, the metaphysical question still stands when dealing with this problem. If causation is taken to be a feature of the world, an aspect of the temporal asymmetry, then connecting it to the manipulationist concept of causation may be problematic.[42] Thus, the best way of making sense of Frisch's considerations and manipulability would be a picture where manipulability-based causal models are useful in virtue of their representation of causation, as one of features of the world. On such model manipulability makes use of causation rather than defining it.

Turning to my project now, there are a few points on which my forthcoming investigation differs from Frisch's. First, Frisch seems to follow Woodward's non-committing realism such that more discussion of the metaphysical foundations is not necessary. As specified in the previous chapter, I believe that looking into causal realism and its connection

[42] Unless one embraces perspectivalism and takes the human perspective to be more fundamental than any of these concepts. However, this answer is not satisfactory for the causal realist.

to manipulability is a worthwhile endeavour, and such endeavour would have more to say about metaphysics. Furthermore, as I hope to show in the next chapters, I do not believe that the distinction between metaphysical and functional aspects of approaches to causation needs to be completely clear-cut. Even if one in not interested in answers to metaphysical questions, there are underlying metaphysical assumptions that affect the workings of a functional project.[43] Secondly, by taking the causal asymmetry to be provided, among other things, by a causal model, Frisch's account might not be satisfactory for someone looking into the source of the causal asymmetry. Causal claims and temporal claims and some counterfactuals may be asymmetric in their causal modelling uses, but the more metaphysically-oriented philosopher might inquire into what renders such models asymmetric. Is there anything more to the causal asymmetry than mere stipulation? This is precisely the question that I am addressing in this chapter. A further interesting question is, in case the causal asymmetry is accounted for, would such account cover the uses of causal concepts outside of causal models (i.e. normally people do not think in terms of causal models when talking about a causal connection between, for instance, throwing a rock in the direction of a window and the window shattering)? Thirdly, Frisch (2013) also argues for a causal theory of time. This is the opposite of the project that I will be arguing for, namely that the direction of causation is provided by the direction of time. Despite the conflicting claims, it should be pointed out that there are some common objections (the issue of backwards causation, time travel, and simultaneous causation) that both projects need to answer and, thus, looking into Frisch's account may be of help in providing answers to these objections.

A final view to discuss on causal asymmetries would be Hausman's (1998) project. I will not go into detail about it, however, since Hausman also takes causal relations and their asymmetric feature as basic. As pointed out by Hitchcock, Hausman may also have to deal

[43] Which I will be illustrating in Chapter 4.

with circularity charges: ‘it is crucial to Hausman's project that the relation of causal connection be conceptually prior to that of causation, so the worry that our best conceptual grip on causal connection will come via the concept of causation needs to be addressed.’ (Hitchcock 2000: 175). While Hausman’s project of identifying the independence relation in most accounts of causation (including Lewis’s counterfactual theory, the agency theory, and manipulability) provides an interesting insight into how causal notions may relate to one another, it does not explain the source of the causal asymmetry. Once again, someone interested in why causal notions come to be asymmetric needs to look for that explanation elsewhere.

2.3 Temporal direction and manipulability

Given a definition of causation as manipulability and a realist perspective on the nature of causal relations, I suggest that the direction of causation can be given in terms of the direction of time. While I will not go for a full account of the metaphysical foundations for manipulability, I hold that the equivalence between the direction of causation and the direction of time can be counted among the features of causal relations as they are in the world. My solution requires that the analysis of causation be broken down into a fundamental level analysis and a higher-level analysis. This kind of distinction has been made by Strevens (2013), through separating two components of a theory of causation: causal influence and difference-making criteria. According to Strevens ‘a theory of causal difference-making, and hence of the truth conditions for causal claims, will have two parts: a criterion for causal influence, determining what relations make up the web of influence, and a criterion for difference-making, determining which aspects of the web make a difference to a given high-level event.’ (Strevens 2013: 309) Regardless how influence may be described, I claim that temporal direction is among the fundamental features that ground causal relations.

In order to illustrate this, I will first make use of some considerations from the previous chapter. As mentioned earlier, according to Woodward's account, metaphysically speaking, intervention is not constitutive of causation. There are, however, some intrinsic features that account for the causal relation as it is in the world, which make it exploitable for the purposes of manipulation and control. My claim is that one of these features, constitutive of causation at a metaphysical level, is temporal direction. I am now going to assume that at the metaphysical level the causal relation can be described in a way that resembles Lewis's influence account.[44] This description need not focus on counterfactual dependence as described by Lewis, but on a kind of dependence that could ground relations related to manipulability such as invariance and sensitivity, mentioned in the previous chapter. These dependence relations ground the higher-level claim that effects can be manipulated through their causes. Without arguing for a specific metaphysics that may ground manipulability, I claim that for the manipulability relation to hold, and for the causal relation to be asymmetric, the more fundamental relation of dependence has to include temporal information such that the entities[45] that ground the variable exerting the influence are temporally prior to the entities that constitute the influenced variable. While I have chosen this description because I believe it to match Woodward's considerations on manipulability, explaining the asymmtery of causation through the asymmetry of time need not be limited to this version. Whatever form the web of influence may take, it has to include temporal information in order for causal relations to be asymmetric.

How does this interact with manipulability as a higher level approach to causation? As specified earlier, a realist account of causation as manipulability admits to the existence of causal relations in the world that ground claims about manipulability. I suggested that

[44] I wish to make it clear that, although I am sympathetic to this particular way of providing metaphysical foundations for manipulability, I will not argue for this approach here. The example is only meant to illustrate my take on causation as manipulability and temporal order.

[45] Such as concrete events (as described by Strevens 2003), or event alterations (Lewis 2000).

conformity with time’s arrow is one of the features of such relations and that controlling effects through controlling their causes makes use of a relation that, among other things, satisfies a temporal order requirement. While it is neither intervention, nor counterfactual dependence that ground the asymmetry of causal relations, their use of an asymmetric relation is grounded in the direction of time. Since, as already pointed out, manipulability needs the direction of causation in order to work, rather than grounding it, there is no problem with claiming that temporal direction is constitutive of causation at a more fundamental level than manipulability. A further question, that I will not be addressing here, concerns the status of the direction of time, namely, whether it should be taken as primitive or analyzed in terms of a different concept. I believe that, as long as the explanation of causal asymmetries goes from the direction of time to the direction of causation one could go for either option when seeking to provide a satisfactory account for the direction of time.

Another potential worry here concerns the relation between the ‘human face’ of causation and the metaphysical foundations. As I hope to have shown by now, it is possible to claim that causation can be defined through manipulability at a higher level while admitting to the existence of some more fundamental features that ground manipulability. The asymmetry of causation is a matter of objective fact insofar as it is part of those fundamental features that manpulability relies on and need not be brought about along with the human perspective. By contrast to the subjective solution advocated by Price and Weslake, on this proposal the causal asymmetry is about the world and not about the subject’s perspective. It is, thus, compatible with causal realism.

I further hold that the connection between causation and temporal direction can also be made as a claim about causal understanding. If one is not convinced by the talk about metaphysical foundations, having more of an interest in the epistemology of causation, one could claim that causal relations can be understood in terms of manipulability while the direction of causation is understood in terms of the temporal direction, since manipulability itself does not suffice for understanding the causal asymmetry. The latter claim draws from some considerations made in experimental context on causality and timely intervention (Lagnado & Sloman 2004). The point here, on which I will expand later on in this chapter, is that people understand the asymmetric nature of causation through connecting it to temporal direction. The connection between causation and manipulability is made at a later time, on the basis of understanding causation and its asymmetric feature in relation to temporal direction. It is also important to note that, in most of the ordinary uses, manipulating causes is done before observing the changes in their effects, so again, the relation satisfies the temporal condition.

One important question regarding my solution is whether claiming that the worldly features grounding the causal relation between two variables, say X and Y should strictly follow the arrow of time, or whether they should simply not go against it. In the former case, the claim that X causes Y involves the claim that Y occurs after X. In the latter case, the claim that X causes Y entails that X does not occur after Y (and thus, they can be simultaneous). While I will be discussing the issue of simultaneous causation into more detail in an upcoming section, I would like to clarify two things right now. One of them is that this problem is very serious for someone endorsing the metaphysical claim but it can easily be solved by someone endorsing only the claim about causal understanding. Someone defending the claim about causal understanding has two main ways of explaining cases of simultaneous causation: a) claim that causal understanding overlaps with temporal direction and that it is more difficult to understand causal structures when the occurrences are simultaneous (there is experimental evidence for this, as I will be pointing out later on) and b) that although the two variables look simultaneous the processes grounding their causal relations occur with some delay that is not noticeable to the observer. Thus, concerning the claim about causal understanding, I hold that typically, causal claims are understood in relation to temporal direction; when cause and effect appear simultaneous, they are more difficult to identify.

Concerning the metaphysical claim, I hold that simultaneous causation is possible at a higher-level perspective on causal claims, and that defending a manipulability view on causal relations helps in singling out the putative cause and effect even if they appear to occur simulataneously. However, I also hold that at the more fundamental level the entities constitutive of the cause variable have some temporal priority over the entities constitutive of the effect variable. The features of the causal relations exploited by an intervention do exhibit a causal asymmetry through the asymmetry of time. Thus, while there may be cases of simultaneous causation, a description at a more fundamental level always involves temporal direction. I will discuss potential objections shortly.

While connecting the direction of causation to the direction of time has at some point been a standard view on the matter (see Hume 1975, orig. 1748; Kant 1965, orig. 1781), more recent approaches, such as those discussed previously, sought different concepts to account for both arrows. I will now go on to exploring some of the main arguments and problems around equating the direction of time with the direction of causation. According to Schaffer's (2014) discussion there are several of them, which I will be discussing in turn.

2.3.1 The argument from bilking and backwards causation

An argument in favour of equating the direction of time with the direction of causation originates in Black's (1956) discussion of backwards causation. Black's argument can count both as an argument for the claim that the direction of causation follows the direction of time as well as an argument against backwards causation.

In Black's original example a clairvoyant is able to accurately tell whether a coin that will be tossed a couple of minutes later is going to land heads or tails. Let us call the two relata in the presupposed backwards causal relation 'A' for the prediction and 'B' for the toss.

Without going into the full details of Black's discussion, his argument against backwards causation, and thus for the claim that the direction of causation follows the direction of time, is that after A occurs, B's occurrence can be prevented or modified. Thus, after the prediction is made, the person in charge with tossing the coin could simply refuse to do it. Also, supposing that the person tossing the coin mastered a technique that would favour one result over the other, could make the coin land heads while the prediction was for tails. This would imply that B is, after all, causally independent from A. As Black points out, claiming that a later event is a sufficient condition for the occurrence of an earlier event and, at the same time, that the latter event is causally independent from the former amounts to a contradiction.

It is interesting to note that there is connection between taking the direction of causation to follow the direction of time and the idea that interventions can be used to change effects through their causes that the argument from bilking relies on. Since my purpose here is to investigate temporal direction as a basis for the asymmetry of causation in the context of a manipulability analysis, the points that I will be making are more or less independent from Black's argument. I would like to point out, however, that an objectivist view on intervention can be used to answer an objection raised by Dummett (1964). Dummett's point is that that the bilking argument does not work in cases where human intervention is impossible, when the event is not repeatable, or when one does not know whether the effect has occurred or not. I will first look at the first case: unmanipulable causes. If one takes a counterfactual concept of intervention (such as the one by Woodward), or explains unmanipulable causes through analogical reasoning (as Menzies and Price do), the bilking argument still stands: one does not need to be able to change the result of the toss, it is enough to formulate the potential result through a counterfactual or through a model that admits human interventions. Another interesting thing to mention is that a similar result may be achieved through Salmon's (1984) earlier concept of causation as mark transmission: one can mark an earlier event and see what happens to a later event, but marking a later event will have no consequence on the earlier event.

Since my interest lies in connecting temporal direction to causal direction within the framework of causal realism, I will argue that Woodward's counterfactual concept of intervention can be used to defend the bilking argument from Dummett's criticism. Relying on Black's original example, even if it is not possible to change the coin toss; it is possible to formulate a counterfactual of the form 'if B had not occurred A would not have occurred'. This counterfactual should hold for most[46] relations where B causes A. In this particular case, however, it does not hold, since we know that A (the prediction) occurs no matter what happens to B (whether the coin lands heads or tails). Now the defendant of backwards causation could make a comeback and say that although, for the most part, there is counterfactual dependence between causes and effects, there are causal relations for which counterfactuals do not hold. Examples of such cases are pre-emption and symmetric overdetermination. Since I am going to rely on a view of counterfactuals and causation similar to Lewis's influence account,[47] pre-emption cases are excluded, since the effect variable would have occurred in a different manner, or at a different time were the pre-empted cause to be activated. This is not the case for backwards causation, where the effect variable (the prediction) can no longer be sensitive to any changes in the cause variable. We are left with symmetrical overdetermination, where there is no counterfactual dependence between cause and effect, since the effect can be brought about at a similar time and in a similar manner by a backup cause. But could one say the same thing about backwards causation? What if after the prediction (A) is made, and the toss is interfered with (B) someone else makes another coin toss (C) that yields into the predicted result? Note that in the case of overdetermination the effect does not counterfactually depend on the actual cause because there is a backup cause.

[46] I will shortly go on to the cases where it might not hold.
[47] See chapter 1.

When we say that there is a causal connection between the actual cause and the effect variable but not counterfactual dependence we rule out counterfactual dependence because of the backup cause. In the case of backwards causation, however, even if there is no counterfactual dependence, can one say that B causes A? No, since it is necessary to have an additional variable, C, to causally explain A's occurrence. Now, since C also takes place after A, a further question is whether C causes A, and whether this can be counted as an instance of backwards causation. Getting back to the analogy with a case of overdetermination, the problem here is that in an overdetermination case if both B and C are causally connected to A, and B's occurrence is prevented, then the counterfactual 'if C had not occurred, A would not have occurred' is true. In the backwards causation case it is false, since, as specified above, A still occurs even if B is prevented and C is interfered with. The upshot is that if backwards causation is to work, then an explanation of why the corresponding counterfactuals are false is necessary. I have looked into an analogy with overdetermination and concluded that the same kind of explanation for counterfactuals not holding is not available for someone defending backwards causation.

I will only briefly discuss the other two cases where Dummett holds that backwards causation may be possible. I believe that unrepeatable events can be accounted for by using counterfactuals as well. For instance, let us suppose that someone claims that World War 1 was not caused by the assassination of Archduke Franz Ferdinand, but by some other event taking place a few hours after the start of World War 1. All of these are historical events and, thus, no longer subject to manipulation. However, based on the knowledge of the other events, one may evaluate the truth value of the counterfactuals (for instance, some features of the putative cause event that fail to be reflected in the start of World War 1). Finally, there are reasons to doubt the claim that not knowing whether the putative effect event has occurred in a case of backwards causation has much bearing on the conceptual possibility of backwards causation. To be sure, it may make sense for someone to keep on performing a certain action that may increase the probability of an effect if one does not know whether that particular effect took place or not (say, keep on taking some medication for an infection that does not present any symptoms without knowing whether the infection is still active or not). However, if causation is taken to be human-independent (as I hope I have shown it to be the case for causal realism in the previous chapter), then the subject's knowledge about an effect having occurred or not should have no bearing on whether there is a genuine causal connection between two events, be it backward or forward.

Reinforcing the bilking argument, along with the further considerations I made about the features of causal relations and how they relate to backwards causation may further raise some scepticism about the possibility of backwards causation. Even if these arguments are not taken to be convincing enough, the main point to make here is that equating the direction of time with the direction of causation does rule out backwards causation.

2.3.2 Time travel

Schaffer (2014) discusses the argument from time travel and ways of answering it. I will not deal with the issue of time travel as such here, but rather with its connections to causality. Time travel seems to be a problem for equating the direction of causation with the direction of time because it involves backwards causation. A strong argument against backwards causation should also rule out the worries about time travel as an objection to an explanation of the direction of causation through temporal direction.

How would time travel work without backwards causation? I will consider an example of a time travel paradox from Futurama, a TV show by D.X. Cohen and M. Groening. In one of the episodes, the main character of the series, Fry, goes back in the past, meets his grandmother and triggers a series of events which lead to Fry becoming his own grandfather. According to the Lewisian (1974) resolution of time travel paradoxes, one could claim that it is possible for Fry to become his grandfather relative to a set of facts containing the past circumstances, his state of mind, his grandmother's state of mind etc. However, relative to another set of facts, containing, for instance, Fry's actual DNA, or blood type, which would be altered were Fry to be his grandfather, Fry cannot become his own grandfather. While I am not going to commit to any particular stance on time travel, I hold that time travel is conceptually possible as long as it does not involve backwards causation. In the particular example I have been discussing, one may ask whether Fry's being thrown back in time caused him to become his own grandfather. Using the same framework, the answer is negative since whatever changes one may make in Fry's trip in the past (for instance, going in the more recent or more distant past, in a specific geographical area or another etc.) the facts related to his parenthood (DNA, blood type etc.) stay the same. However, while looking at the sequence of the past where Fry ends up and meets his grandmother, one could make a set of true causal claims, for instance that Fry's attempt to keep his grandfather away from harm causes his grandfather's death, which in turn determines further interactions between Fry and his grandmother. However, none of these causal interactions are to be taken into account when looking at the set of facts from a global perspective, where one could inquire whether some future event, such as the event that triggered Fry's being thrown back in time, could cause changes in the past, such as Fry's parenthood. Thus, in the case of time travel, only forward looking causal claims can be true. These considerations overlap with the claim that time travel is possible only though locally forward steps.

2.3.4 Simultaneous causation

I take the simultaneous causation problem to be the main issue for the metaphysical formulation of my proposal. One thing to note in relation to the previously discussed objections is that someone accepting the possibility of simultaneous causation can also reject backwards causation, as such a stance would concede that causes can precede or occur at the same time as their effects, but they cannot happen after the effect has occurred.

There are several stances on the possibility of simultaneous causation. For instance, Mellor (1995) argues that the examples in the literature do not make a convincing case for simultaneous causation, but the conceptual issue is nevertheless worth investigating. On the other hand, the literature on powers and causation holds that simultaneous causation is possible (see Marmadoro), or even that causes and effects are always simultaneous (Mumford and Anjum). I will go on to discussing some of their arguments in what follows.

Mumford and Anjum (2011) argue for a theory of causation based on powers. I will not go into details about the characteristics of the theory, as I will only be interested in the claim that causes and effects are always simultaneous. Mumford and Anjum define causation as properties producing effects in virtue of their powers or dispositions. In Glynn's (2012) reconstruction, the authors make an analogy between powers and forces and represent them as vectors. Causal powers can be represented as vectors going into several directions in a quality space with a resultant: 'given a number of powers at work in a causal situation, the powers – like forces - may combine in an additive way so that the situation overall disposes with a certain magnitude in one direction or the other' (Glynn 2012: 1101) Given that the cause variable is given by an exercising of powers and the effect variable by their manifestations, and that the authors claim these to be simultaneous, causation is always simultaneous.

It is obvious how such claim would go against the idea that the asymmetry of causation is grounded in the asymmetry of time. I believe that Glynn's reply is convincing enough as to shed some serious doubt on the claim that all causal relations are simultaneous. Glynn's (1102-1103) counterargument brings fourth relativistic physics, where absolute simultaneity is impossible. As Glynn points out, 'if events (or property-instances) c and e are simultaneous relative to any frame whatsoever, then unless c and e occupy precisely the same space-time location, the two events lie outside of each other's future light-cones.' (Glynn 2012:1102-1103) The latter possibility would violate the requirement that causal processes not propagate faster than light. It is hardly the case, however, for most or all instances of causation cause and effect occupy the same point in space, so causation cannot always be simultaneous. I will postpone the discussion of causation being sometimes simultaneous until after discussing another case for simultaneous causation.

Another field where simultaneous causation is important is the Aristotelian view on powers, as supported by Marmodoro (2009) and her collaborators. According to Marmodoro's (2013) interpretation, Aristotle defended a theory of causation as ontological dependence. Once again, I will not go into the details about the theory, but discuss the way in which it endorses simultaneous causation. When discussing the direction of causation the following claim is made regarding to Aristotle's work:

> Contrary to what common sense might lead one to think, Aristotle is clear that actual causes do not precede their actual effect in time; teaching and learning (by being taught) have the same life span. The potential to teach is in the teacher before she engages in actual teaching (...) and so is the corresponding passive power in the learner, but their actualization is one and the same (hence there is complete overlapping in time) (Marmodoro 2013: 15)

According to Marmodoro, despite the temporal overlap, causation takes a direction as only the patient changes through transmission of the form/cause, and not the agent.

It is clear that someone opting for an Aristotelian view on powers (as in Marmodoro 2009) and for a theory about causation based on powers would stipulate that cases of simultaneous causation are at least sometimes possible. Thus, even if there may be convincing arguments against the claim that causation is always simultaneous, someone accounting for the asymmetry of causation in terms of the temporal asymmetry needs to handle these cases.

My reply to such cases is to point out that cases of simultaneous causation only apply to higher level causal claims. While I do not wish to go into the details of theories defining causation through powers, some details of the Mumford and Anjum account will be relevant here. Mumford and Anjum define causation in terms of powers, but at the same time do not reduce powers to their categorical bases or to a conditional analysis. Furthermore, dispositions are causal notions themselves. Thus, on this line of thinking, there is no possibility for a noncircular analysis of causation in terms of powers of dispositions. As specified earlier, this need not be a drawback; the theory can still be informative. However, this also means that causation, and specifically its direction, may be described in a noncircular manner at a more fundamental level. Going back to Mackie's argument above, the question is whether claiming that the cause variable has the power to bring about the effect (say throwing a vase against a hard surface can make it break) while this power is not exhibited by the effect variable. On a first glance, this solution may explain the direction of causation without resorting to temporal direction. However, as mentioned earlier, in the Mumford-Anjum theory, dispositions are causal notions. Thus, once again, the issue is that in order to explain why the causal connection between dropping a vase and its subsequent break is asymmetric, one needs to employ other causal connections (those causal claims that serve as grounds for fragility). The problem of the direction of causation is still left unanswered.

However, if my earlier suggestion is right, and temporal direction can be counted among the features constituting causal relations at a more fundamental level, then it is possible to account for the asymmetry of causation in terms of the asymmetry of time and also have higher level causal claims that admit of simultaneous causation. The connection between causation and powers would have to be made at a higher level with causal relations being constituted by other features in the world. These considerations may not apply to a view on powers such as the one by Marmadoro (2009), where other powers are not necessary for accounting for powers, but that would require full blown theory of causation based on an Aristotelian view on powers. Finally, I believe that the Aristotelian claim that cause and effect are simultaneous can also be subject to controversy. Looking at the teaching and learning example from the perspective that I am proposing, one could break down the two causal relata. For instance, in a philosophy class, if the student is listening to the teacher enouncing an argument, the student will have to listen to every premise and the conclusion before learning the argument. Thus if premise 1 is enounced at t0, premise 2 at t1 and the conclusion at t2, one can say that the student learns or understands the argument at t2, and thus, there is a temporal delay between the start of the teaching activity and the realization of the learning activity. Even if one were to claim that the student might learn the argument along with the last premise (t1), that would still be at a later time than the statement of premise 1 (t0). Thus, I believe that the teaching/learning example can be spelled out as a case of simultaneous causation as a higher level claim, but it does involve temporal direction when analyzed at a more fundamental level.

Regardless of whether my solution works at the metaphysical level, I believe that this problem can be answered from the perspective of a weaker claim linking manipulability to time. While I will discuss this claim at the end of this chapter, I will provide a sketch of the answer here. Even admitting that from a metaphysical point of view simultaneous causation may be possible, in everyday contexts causal relations are linked to judgments about temporal direction, and thus, people find it easier to recognize a causal relation when presented with temporal evidence that when presented with simultaneous occurrences of cause and effect.[48] Furthermore, this feature can be of use in scientific contexts where one is looking for a specific cause and on the basis of temporal cues events happening after the putative effect are ruled out.

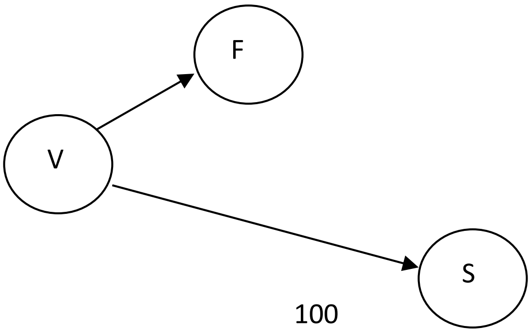
[48] Again, I will have more to say about the empirical aspect of this in chapter 4.

2.3.5 Physics

This objection is to some extent related to the previous considerations on backwards causation. The worry is that at some point there have been theories in physics that posited backwards causation. As Schaffer (2014) summarizes the discussion, there are several ways of responding to this objection. One is to consider that such theories are false, or that they will turn out to be false at some point. Another answer is that even though such theories might in fact be true, there may be no account of the direction of causation compatible with their claims about backwards causation. Finally, there are philosophers who take the symmetries from the laws of physics to imply that there are no asymmetries in the world, and the direction of causation is a result of people's projections as agents (as in the Price-Weslake account discussed above). On the other end of the spectrum, there are philosophers who try to ground the causal asymmetry in physics (see Albert and Loewer discussed above). As pointed out in the previous section, the latter tries to explain both the direction of time and causation through a third arrow. While I believe this to be a promising way to go for explaining causal order, besides dealing with counterexamples, it would still need to build a full account on causation based on the considerations on the second law of thermodynamics and the past hypothesis.

2.3.6 Joint effects

I have previously discussed the problem of joint effects by reference to different accounts of causation. I will thus, make use of the same example here of a virus (V) present at time t0 that causes fever (F) at t1 and spots on the skin (S) at t2.



While F and S are temporally ordered, F should not be counted as the cause of S. While there are several ways of answering this problem, I wish to point out that my use of temporal direction is meant to capture only one of the features of causal relations, namely the asymmetry feature. Thus, I do not seek to define causality exclusively in terms of temporal direction. There are other features that allow for such relations to be used for the purposes of manipulation and control. Since S cannot be manipulated through F, it does not count as its cause.[49]

2.3.7 Causal theory of time

Some philosophers reject the option of explaining the order of causation through the order of time because it would entail the impossibility of a causal theory of time (or at least a reductive version of such theory). As noted in Schaffer (2014), such an option would also have the previously presented issues with time travel or physics, and simultaneous causation. While the standard answer amounts to rejecting the causal theory of time, I will not go into that debate here, and acknowledge that this way of accounting for the causal asymmetry rules out the possibility of a reductive analysis of time in terms of causal relations.

2.4 Temporal direction and causal understanding

To sum up the discussion of the problems and objections, I would like to emphasize that my interest lies partly in manipulability along with a realist view on causation, and partly in the ways in which this approach can be applied to causal claims in psychological or more general scientific contexts. Thus, while I acknowledge that adding a temporal component to the metaphysical foundations of manipulability may rule out some philosophical analyses of certain concepts, I do not see these objections to be decisive, at least as long as the more pragmatic perspective is taken into consideration. As pointed out above, what I consider to be

[49] Unless F is manipulated though V. On both Woodward's concept of intervention and Lewis's counterfactual analysis that option is excluded.

the most serious issue that remains is the possibility of simultaneous causation. I will point out below how it can be dealt with from the perspective of causal understanding.

Having presented the metaphysical perspective on causation and time, the question that may be asked here is whether temporal direction should be taken to ground the asymmetry of causation at a metaphysical level, or whether it could also be linked to causal understanding, in the sense that people take causal relations to be asymmetric because causal judgments are linked to temporal judgments. Before making some general considerations on the two viewpoints, I will present the problem of the causal asymmetry through the perspective of causal understanding.

Generally, causes are thought of as preceding their effects, and one expects to bring about an effect through bringing about its cause, but not the other way around. What is the source of this feature of causal thinking? One answer is that causal judgments are closely related to temporal judgments. Claiming that X causes Y usually involves a temporal claim as well, that X precedes Y, or that Y does not succeed X. Seeing several instances of X occurring and being followed by Y often leads to conjecturing a causal connection between X and Y. Temporal order is not a sufficient condition for a causal connection, however: X and Y may just be randomly correlated, or they may be results of a common cause. Thus, the relation between X and Y can be singled out as causal through other means, such as a counterfactual test, an intervention, or information about a mechanism linking X and Y. While causal connections involve a more complex set of conditions and tests, my claim is that temporal order is sufficient to account for one particular feature of causal relations: asymmetry. Note that I am making this proposal here from a purely epistemic perspective; I will present some empirical evidence from causal learning in chapter 4. How does that deal with the problems discussed above? By moving the issue of causal order from the metaphysical to the epistemic realm no claims are made about the nature of time or causation. Thus, one would be able to advocate for the possibility of a causal theory of time, or backwards causation, or simultaneous causation. Under my proposal, such concepts would be more difficult to grasp by causal reasoners. Once again, there is empirical evidence that I will be discussing in chapters 3 and 4 in the sense that people find it easier to infer from cause to effect that the other way around, or to infer a causal relation from temporal cues than from cases where the causal variables are simultaneously activated.

I would now like to move on to the interactions between the two claims. I have been trying to show that endorsing the claim about causal understanding does not necessarily lead one to particular conclusions about the metaphysics of causation. While temporal order may be a useful tool for thinking about causal order, there may be some other feature in the world accounting for the asymmetry of causal relations. I would also like to point out that such a view would also be compatible with a projectivist or fully epistemic view. Since the options are pretty much open, these considerations are in line with Woodward's view on metaphysical foundations. However, since I have expressed my interest in the realist version of manipulability, these options do not seem to shed much light on the relation between the causal asymmetry, causal relations as they are in the world, and their support of manipulation and control. I would thus like to point out that taking the direction of causation to be connected to the direction of time at a metaphysical level may explain why people are successful in their use of temporal cues when inferring causal structures and why they generally think of causation as an asymmetric relation. This way of thinking is analogous to Woodward's argument for the connection between worldly features of causal relations and successful causal inference presented in chapter 1.

Looking at the other way in which the distinction cuts, another question to ask is what would accepting the metaphysical claim entail. That is, would it entail anything about the way in which people reason about causes? I believe that an important thing to point out here is that while having temporal order play an important role in both the metaphysical and causal understanding case may provide an explanation for success in causal inference, it does not make this way of thinking about causal relations the only valid one. As mentioned above, there are several tests that can be used to single out a relation between variables as causal. Even if at a metaphysical level the order of causation may be given by the order of time, the understanding of causal relations as asymmetric can be grounded in one or several concepts that would make causal claims asymmetric. While at a metaphysical level reduction is an issue, at the level of causal understanding the issue is to a large extent empirical, namely, how people learn that causes have a kind of priority over their effects. Another interesting thing to note is that, opting for the manipulationist test for identifying causal relations along with the temporal account for the causal asymmetry at a more fundamental level yields into the consequence that both facts about causation and facts about manipulability are influenced by the temporal situation of the agent. This may be seen as a way of turning the Price-Weslake picture on its head: the temporal direction is taken to be an objective feature of the world, experienced by every agent and thus affecting the agents' understanding of causal concepts as opposed to being a projection of the agent's deliberative situation onto the world.

While, as pointed out above, there are problems with the metaphysical perspective on time and causation, and also, the step from the metaphysical to the causal understanding perspective may not be necessarily warranted, I believe that bringing the two claims together would result in a coherent, general perspective on causation and causal inference. I also take this to be the way to go for someone interested in both the uses of manipulability based causal claims and causal realism.

Finally, at the end of the previous chapter I was discussing Price's objection to causal realism. To recall, the issue was whether a manipulationist concept of causation can be extended to a universe where two spatio-temporal parts with opposite temporal parity are connected by two wormholes. The causal projectivist's answer is that causation is dependent on the position of the agent (next to one of the two worm holes). Given the previous discussion, I claim that the causal realist may answer that a manipulationist concept of causation can be extended to the area of the universe where time runs in the opposite direction. However, causation need not be dependent on the position of the agent, but on the nature of time. Thus, what causes what depends, among other things, on the way in which the time runs; and this fact can be grounded in the structure of the universe (arguably, different from what the structure of the universe is taken to be right now and consequently, our concept of causation would have to change, following either one temporal arrow or the other).

I will end my discussion of metaphysical foundations for manipulability here and go on to investigating how the connection between causal understanding and temporal order may help the manipulability view to make sense of some experimental data. I should note, however, that I will keep on discussing the way in which metaphysical assumptions about causal relations play a role in the investigations on causal reasoning. Thus, even though I will be focusing on the pragmatic aspect of manipulability, I will do so from a more general perspective, where metaphysical and pragmatic issues are intertwined. In the next chapter I will move to the functional approach to causation that Woodward presents with the aim of investigating later on how manipulability is applicable to causal reasoning tasks.

Chapter 3: Causation as a Functional Concept

In this chapter I will be concerned with the manipulability account as a functional account of causation, as presented in Woodward (2014). I will connect the functional way of describing manipulability to the asymmetry issue discussed in the previous chapter, and I will go on to pointing out some further issues. The main problem that I wish to raise is that a view that emphasizes the usefulness of causal judgments might expand beyond the features of the interventionist account. I will point out how some of the features of Woodward's manipulability view may constrain the area of applicability of the interventionist concept of causation. I will also take another look at the suggestion of linking manipulability to temporal order, discussed in the previous chapter, and explain how this suggestion would be of help in overcoming some of the constraints. These considerations are meant to clear the ground for an investigation of the application of manipulability as a functional account of causation to psychological work on causal reasoning and causal learning. While this will be the main task for chapter 4, I will make some distinctions and delimitations here.

3.1 Woodward's functional approach to causation

Woodward (2014) sheds some light on the aims and goals of his theory of causation as manipulability that have been previously subject to debate. Delimiting his project from what he deems the metaphysical, the descriptive, and the fit with physics projects, Woodward describes his account as follows: 'by a functional approach to causation, I have in mind an approach that takes as its point of departure the idea that causal information and reasoning is sometimes useful or functional in the sense of serving various goals and purposes that we have. (...) 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

2014: 693-694) This approach is consistent with some features of the manipulability account (like the emphasis on methodology, or the circular definition of intervention) discussed previously, and it also enables answering certain questions that may or may not constitute the focus of the other projects. Two of the questions brought about by Woodward will be important for my purposes here. One of them concerns the procedures of testing causal claims. The other concerns the way in which the normative theory endorsed by Woodward's functional account relates to descriptive investigations into causation in psychological contexts. I will go into more detail about these issues in chapter 4, when I will also discuss examples.

An issue that I briefly discussed in chapter 1 and to which I would like to come back right now is the idea that the functional account requires no metaphysical commitments beyond a claim to modest realism. While I find the link between the philosophy of causation and actual uses of causal claims that the functional account draws to be an interesting way to go when investigating the problem of causation, I do not agree with a clear-cut delimitation from metaphysics. For one, as I will be arguing in this chapter and the next one, the interventionist account has some features, other than the commitment to realism, that could be explained through some assumptions connected to the metaphysical debate around causation. One of them is the idea that facts about manipulability may be more fundamental than facts about mechanisms and they could consequently be explanatory of causal mechanisms. The other is that in the case of causation the standard interpretation of counterfactuals should rule out backtracking, mainly for the purpose of keeping causal relations asymmetric.[50] These features may pose problems given that manipulability (or Woodward's concept of manipulability, in particular) is not the only concept of causation that applies in functional

[50] It may still be subject to debate whether this is a metaphysical problem or merely a problem about the semantics of counterfactuals. Since I believe that there is an important link between the non-backtracking interpretation of counterfactuals and the asymmetry of causation, I also think that this interpretation is, at least partly, linked to a metaphysical issue. This point is enforced by the fact that Woodward admits to causal and merely correlational relations as being separated on metaphysical grounds.

context. These observations, however, merely show that Woodward's version of a functional account of causation is hardly metaphysics-free. Could a stronger claim be made, namely that there is no precise place where to draw a distinction between a purely metaphysical and a purely functional account of causation?

A first thing to look at when tackling this question concerns the nature of the distinction. Since I will not be so interested in the fit with physics project, I will base my discussion here on the distinction between metaphysical, descriptive and functional projects. While Woodward acknowledges that the distinction is neither exclusive nor exhaustive, he opposes the functional project to the descriptive and metaphysical projects. Thus, a question to ask is what separates these projects.

One thing to note is that the above-mentioned projects of investigating causation employ roughly similar methods: conceptual analysis along with some formal structure or logic that is characteristic of philosophical arguments. One may add an empirical side to the functional account, but I believe that it can be argued that, in the way Woodward describes it, the descriptive project could also benefit from studies into how people reason causally. Also, as I will note shortly, some of the 'Canberra plan' philosophers took ordinary assumptions about causation to act as constraints on philosophical approaches to causation. If these ordinary assuptions are studied in an empirical manner, then they are relevant for a descriptive as well as for a functional project. Furthermore, since Woodward includes philosophers such as Lewis or Menzies in the descriptive camp, it can definitely be claimed that the aim of the descriptive project has a strong metaphysical component. Thus, given that these projects seem to be concerned with the same general topic (causation) and employ similar methods (the way in which philosophical investigations work and maybe some empirical work into causal reasoming), the delimitation cannot be made on these grounds. But what is making the delimitation?

My interpretation is that what makes these projects distinct is a difference in focus: the metaphysical project is concerned with the facts grounding causal claims, the descriptive project tries to build a theory of causation that matches everyday intuitions about causal claims, whereas the functional project emphasizes the usefulness of causal claims based on a certain concept or approach. While this difference in focus brings about a stronger preocuppation for some particular topics (i.e. in the case of the functional project how is a such-and-such analysis of causation useful for answering a certain problem), that does not exclude topics that may have more relevance for the other projects. Thus, in the case of the functional account one could ask why does a manipulability-based concept of causation work and why the corresponding causal claims come to be true. An answer to this question will need to make some metaphysical claims, especially if the explanation goes along the lines of causal realism. While I do not claim that a functional account, and Woodward's account in particular, ought to say something more about metaphysics, I believe that a more general perspective on causation in philosophy as well as in other fields relying on causal claims would include some considerations about metaphysics. Furthermore, as I mentioned earlier, and as I will be discussing in section 3.4, I do not believe that, as far as metaphysical problems are concerned, the line between a metaphysical and a functional account of causation is drawn right after the realist claims. There are other features to the interventionist account that can be traced to specific metaphysical stances.

One more thing that I would like to clarify here is that the view that I am defending here does not collapse into what Woodward takes to be the 'turning every empirical issue into metaphysics'. I agree with Woodward that is is useful to look into how causal claims work in functional contexts, and that may not necessarily bring about any metaphysical claims, but I believe that when considering the bigger picture metaphysics comes back into place. Rather than taking this to be an undesirable consequence, I take it to provide a good reason for conducting more research into the connection between manipulability, causal realism and their adequacy to a functional project. Furthermore, I hold that a functional account of causation should have a degree of generality such as to be able to account for the various ways in which causal claims work. From this perspective, interventionism by itself seems to be insufficient.

A final thing about Woodward's description of the functional account that I would like to discuss is the claim that the interventionist view on causal relations need not apply to every domain, as Woodward puts it '"Doesn't work everywhere" does not imply "Works nowhere"' (Woodward 2014: 702). An important domain where causality cannot be thought of in terms of interventions in Woodward's sense is fundamental physics.[51] An observation that I would like to make at this point concerns the pluralist idea that different concepts of causation may work better for different domains of inquiry. Although Woodward does not endorse it explicitly, it seems to be implied in his discussion on causation as a functional concept. It is interesting to note that this parallels his considerations on a difference-making and geometrical-mechanical concept of causation in psychological contexts, where he contends that there could be contexts where one of the two would work better. Again, I will have more to say on this issue, namely even if one were to support pluralism, the question concerning what should be the relation between the two and what is the status of manipulability is still open.

To sum it up, the idea of a functional account of causation seeks to shift the focus from the issues traditionally associated with the metaphysics of causation[52] to an inquiry into the usefulness of an interventionist concept of causation in scientific context, or in dealing with problems from different areas of philosophy (such as the exclusion problem discussed in Woodward 2014). An interesting question to raise here is whether a related project may be

[51] See Woodward 2007b.

[52] Although a notable exception, not discussed in this article, but obvious from Woodward (2003) is the avoidance of counterexamples.

present in the causation literature. Although the discussion differs in some respects, there seem to be important similarities between what Woodward is trying to do with his functional approach to causation and the so-called Canberra plan. I will look into it through some work by Menzies.

3.2 The functional account of causation and the Canberra plan

Menzies (1996) employs Lewis's (1972) strategy for analyzing theoretical terms in order to provide an account of causation. The idea is to explain the meaning of a term through the role it plays in folk psychology. As Menzies puts it 'one can think of folk psychology as a kind of theory consisting of platitudes that are common, albeit tacit, knowledge among us.' (Menzies 1996: 97) In the 1996 article, the platitudes identified by Menzies are:

1. Causation holds between distinct events.

2. The causal relation is an intrinsic relation between events.

3. A causal relation is linked to the relation between two distinct events, where an event increases the chance of the other. (after Menzies 1996, section 3)

While there may be counterexamples to the last platitude, it captures the way in which typically the concept of causation works in folk psychology.

In a later article, Menzies (2009) abides by the same view, although acknowledging that 'it is not necessary (...) to hold this view to think that the platitudes that ordinary people associate with causation should play an important role in philosophical discussions about causation' (Menzies 2009: 341). Such platitudes, Menzies holds, should act as constraints on a theory of causation. I will not go into the details of Menzies' account on the platitudes connecting to contrastive, normative and context-sensitive aspects of causation. However, it will be relevant for my discussion here to point to Menzies' criticism of the platitude that causation is a natural relation between events. His criticism, based on considerations on extensionality, event fragility and absences is still part of the Canberra plan strategy: ‘this is not a rejection of the method of appealing to platitudes, but rather a plea to reject the incorporation of unquestioned metaphysical assumptions into this method’. (Menzies 2009: 343)

Having briefly described the way in which the proponents of the Canberra plan discuss causation, I would like to point to a few similarities and differences in comparison with Woodward’s functional approach to causation. A first thing the two approaches have in common is acknowledging that causation and causal concepts play a role in everyday reasoning. The way in which causal relations are understood in such contexts should impose certain constraints on how philosophers should define causation. It should be pointed out, though, that Woodward does delimit the functional project from what he deems the descriptive project.[53] That may be explained, in part, by the fact that Woodward’s concept of causation is not meant to deal with everyday uses, but rather on how people ought to reason causally, and also with causal explanation in science. Thus, while both projects emphasize the usefulness of causal claims, the aim of the functional project is broader than an analysis of causation through its role in folk psychology. Another way of understanding the delimitation from the descriptive project is through the normative dimension of Woodward’s account. The interventionist take on causation is supposed to say how people ought to reason about causes. In this sense, it seems to be the reverse of the Canberra plan, where everyday intuitions about causal relations were seen as constraints for a philosophical account of causation. Nevertheless, this difference does not run so deep, as Woodward points to the connection between descriptive and normative aspects of reasoning about causation.[54]

[53] According to Woodward (2014), the descriptive project involves fitting ordinary intuitions about causation with a philosophical approach. His examples include more metaphysically oriented approaches than the Canberra plan, such as the one by Lewis.

[54] See, for instance, ‘we should take seriously the possibility that people’s causal cognition is often fairly well adapted to the problems they face or goals they are pursuing.’ (Woodward 2014: 703)

Another interesting aspect to discuss here concerns metaphysical assumptions. While dropping questionable metaphysical assumptions would be very much in line with Woodward's refusal of doing metaphysics, I believe that his functional account relies on the platitude that Menzies argues against (i.e. that causation is a natural relation). One proof for this lies in the allegiance to realism, and in connecting the success on causal tasks to worldly features about causal relations. Once again, I take this contrast to be an example of the fact that, even though none of the authors endorses a fully metaphysical view on causation, disagreements stem from their tacit endorsement of a realist or projectivist view on the nature of causal relations. Since denying the claim that causation is a natural relation is not a core feature of the Canberra plan, but a claim made by Menzies, it does not have much bearing on the relation between the two projects. Nevertheless, I believe it is worthwhile to single out some of the metaphysical assumptions that motivate approving or rejecting certain platitudes.

An issue closely related to the point made above concerns the objectivity of causal relations, and, by extension of Woodward's concept of manipulability. As specified previously, this is the main point where metaphysical issues come in. The question that I wish to raise is whether the objectivity of causal relations can be taken as a platitude, rather than a claim about metaphysics. As emphasized by Woodward (2003), it is important that causal relations stay the same regardless on the changes in human agents' capacities. This requirement can be understood at a conceptual level, as in how we think about causal relations in general, but also at a methodological level, as in how controlled experiments are done and how objective causal relations have to be in order to obtain reliable results from the experiments. It is, thus, important to emphasize the practical advantages of an objective take on causation. Nevertheless, that also brings about a commitment to causal realism and new objections to answer.[55] I believe, thus, in conformity with what I have been stating before

[55] See, for instance, Price's (2014) sceptical challenge against causal realism discussed in chapter 1.

about the mix of epistemic as well as metaphysical issues around causation, that the issue of causal realism and objectivity has both a metaphysical as well as a functional side. The former concerns what causal relations out in the world are like, and how they ground the truth of causal claims based on interventions. The latter concerns the ways in which it is useful to have a concept of causation that does not depend on the agent's goals and capacities.

Coming back to the two projects, another important difference to emphasize is the treatment of counterexamples. The Canberra plan admits that some platitudes or conjunctions of platitudes might have some unresolved problems, but holds that in an analysis of causation as a folk concept it is important that such platitudes hold for most (but not necessarily all) cases. While there is no discussion of counterexamples in Woodward (2014), it is quite clear from his earlier, lengthy, analysis (Woodward 2003) of the advantages that the manipulability theory offers over related approaches in terms of dealing with counterexamples that this is a point where the two projects diverge. Nevertheless, moving from counterexamples to the areas of applicability of the manipulability theory, it seems that both the functional project and the Canberra plan do not require their descriptions of causation to work everywhere. As I have mentioned out above, Woodward (2014) does not require that the manipulationist concept of causation work everywhere.

Another important difference concerns the relation between causation as a folk concept and psychological issues surrounding causality. While according to the Canberra plan, causation is linked to everyday reasoning, no empirical work is brought into the discussion. Woodward, on the other hand, points to experimental work that validates some of his functional claims about causation. Furthermore, he takes some of the theoretical work on the connection between intervention and causal understanding to provide psychologists with new ways of designing experiments. Once again, I believe that this link may be explained through the emphasis on the philosophy of science side of the concept of causation as manipulability.

A final thing that I would like to mention here, and that I will discuss into more detail in the section 3.4, concerns the status of certain platitudes in the context of the overall functional project, or folk psychology respectively. While the Canberra plan admits of any conjunction of platitudes that yields into a satisfactory account of causation as a folk concept Woodward's functional account is centred on the notion of intervention. As mentioned before, it is a very specific concept of intervention, one that fits in with a directed acyclic graph approach to causation. While the particularities of interventionism are of help in dealing with counterexamples, or describing certain common practices in science, as I will be arguing later on, applying them to everyday cases of causal reasoning can raise certain difficulties.

3.2 The asymmetry of causation revisited

I will now have another look at the problem of the asymmetry of causation, as well as to the solution proposed in the previous chapter in the light of the manipulability theory as a functional account of causation. In the previous chapter I have relied on an argument by Mackie's in order to show that manipulability has difficulties in explaining in a noncircular manner why causal relations are asymmetric, or how people come to know that they have this feature. While I have been emphasizing the metaphysical aspect previously, the question to ask now is how the asymmetry of causal relations can be characterized from a functional point of view. I have mentioned that there is no explicit answer to the question regarding the grounds for the causal asymmetry to be found within Woodward's account. Is that a problem from a functional point of view, does it serve our goals to think of causal relations as asymmetric, or could they work both ways? From the point of view of the manipulability theory the fact that causes come before their effects is more or less taken for granted. In chapter 2, I proposed explaining the causal asymmetry through the temporal asymmetry.

From an epistemic point of view, the connection can be explained by pointing to a link between causal and temporal judgments. In this section I will point to some empirical evidence that, ordinarily, people find it easier to reason from cause to effect than the other way around. I will use these findings to justify the claim that even if one were to subscribe to a functional account, one needs to explain the asymmetric feature of causal relations.

In a review study by Sloman & Lagnado (2015), several inquiries into comparing inferences from cause to effect to inferences from effect to cause are discussed. The authors point out that if causal relations are to be thought of exclusively in probabilistic terms, then people should have no problem inferring from cause to effect, as well as from effect to cause. Nevertheless, there is evidence that people find it easier to make a causal judgment as opposed to a diagnostic judgment. To use some evidence cited by the authors, in a Fenker et al. (2005) study, people were better at identifying a causal relation when presented with cause to effect evidence, rather than the other way around. In another study by Hong et al. (2005) children were better at drawing cause to effect inferences than effect to cause ones. A Tversky and Kahnerman (1974) study shows that people assess the probability of effects given their causes higher than the probability of causes given their effects even when no significant difference between the two probabilities is to be expected. In the same study, Tversky and Kahnerman suggest that people are able to run simulations that go forward in time, while diagnostic reasoning would require a different ability. The conclusion drawn by Sloman and Lagnado is that mental simulations require more information than probabilistic dependency: they require a temporal order (see Schwartz & Black 1999) and an understanding of the mechanism through which causes bring about their effects. While I will come back to these two components shortly, I will now address the question concerning how this connects to manipulability and to the more general way in which a functional account of causation accounts for causal asymmetries.

The psychological data on the asymmetry in causal reasoning shows that the grounds for taking causal relations to be asymmetric are not exclusively metaphysical. Thus, a functional account of causation should take this aspect into consideration. The question is: can a satisfactory explanation (of the general idea that causes bring about their effects and not the other way around, or of the discrepancies between causal and diagnostic reasoning) be found within the manipulability account? While judgments about manipulability seem to be closely connected to causal judgments, at least in Woodward's version of manipulability there seems to be no information about temporal order or mental simulation. Thus, at least according to psychological work on this matter so far, interventions play no role in the fact that people find it easier to reason from causes to effects rather than the other way around. An important question to ask here is whether it should. After all, the functional account does not require the interventionist approach to causation to work everywhere. I will address this question at the end of next section after discussing the distinction between the interventionist concept of causation and the geometrical-mechanical one.

3.3 Two concepts of causation and the problem of applicability

In this section, I would like to address two different issues that will be of help in spelling out some limitations to the extent to which the interventionist concept of causation may be applied in causal contexts. The first issue concerns the distinction between difference-making and geometrical-mechanical concepts of causation. This distinction is discussed in psychological context in Woodward (2011a). Rougly speaking, the difference-making concept of causation is based on the idea that causes make a difference to the occurence of their effects. The geometrical-mechanical concept of causation is based on the idea that causes somehow bring about their effects (through a process, transfer, or mechanism that goes from cause to effect). The origins of the distinction can be traced to Hall's (2004) delimitation of two concepts of causation. It is important to point out that the motivation of having such a distinction in the first place was mainly metaphysical, namely to show that the set of requirements that a theory of causation is supposed to meet were only met jointly by the two concepts of causation as production (what Woodward calls the geometrical-mechanical concept) and dependency (what Woodward calls the difference-making concept). Another thing to mention is that difference-making can admit of several criteria. Thus, a counterfactual test, an intervention, or statistical information could all be used as indicators that a putative cause variable makes a difference to the occurrence of an effect variable. The reason why I am mentioning this is that even if one could contend that a causal claim is based on a difference-making account, the truth of the respective claim could be warranted by either one of the criteria. While in a psychological context Woodward's discussion of difference-making focuses on interventions, it is important to keep in mind that the difference-making concept of causation could also admit of different ways of inferring causal structures.

A thing to clarify here is the relation between these two concepts. There are several differences between the two. I mention the most important ones, then discuss the most relevant one for the point that I am trying to make. The differences highlighted in Woodward 2011a are:

1. Area of applicability. Whereas the difference-making concept of causation holds for all sorts of causal relations, the geometrical-mechanical one seems to hold only for physical entities or cases where some process or energy transfer is involved.
2. The difference-making concept of causation sees causal relations as comparative (thus, considering alternative scenarios where the putative cause variable might have been different). This is not the case for the geometrical-mechanical concept.
3. Generally, the difference-making concept of causation holds for type-level causal claims, whereas the geometrical-mechanical concept holds for token-level causal claims, or actual causation.

4. The features that each of the two concepts possesses are neither necessary, nor sufficient for the presence of the features of the other concept.

I would now like to make some observations on the first point above, that of the area of applicability for the interventionist difference-making criterion. While Woodward points out an area where the difference-making concept is more general than the geometrical-mechanical one (causation between mental entities), there are also areas that the geometrical-mechanical concept covers better than the interventionist concept. These cases, as I will point out shortly, include the asymmetry in causal reasoning and the use of temporal cues in identifying causal relations. A further point that may be relevant for the issue concerning the area of applicability of the two concepts of causation may be taken from the debate with Waskan on counterfactual versus mechanistic models of scientific explanation. Waskan (2011) takes Woodward's account to attempt at marginalizing and assimilating the virtues of the mechanistic approach to causal explanation, while Woodward argues that the geometrical-mechanical view on causation cannot be taken to be more fundamental than the manipulability one. I will come back to this debate in chapter 4.

I would now like to go through the claims made above about how widely applicable Woodward's interventionist concept of causation is. As mentioned before, Woodward does not discuss the asymmetry issue. While it is obvious that in order for the interventionist concept of causation to work the way it does, causal relations need to be asymmetric, there is no account explaining why it is so. By contrast, a geometrical-mechanical concept of causation takes the idea that causes and effects are spatiotemporally contiguous to be an essential feature of causation. Applying the main elements of the geometrical-mechanical approach to psychological issues about causal reasoning and inference can yield into an explanation of how people find out that causal relations are asymmetric: once they grasp the mechanism information, they can use it for mental simulation. The interventionist concept, on the other hand, contains no temporal information that a mental simulation for causal inference would require. The other, related, issue is that temporal cues are taken to be indicators of causal relations. Once again, this connects to the geometrical-mechanical concept according to which causal judgments include temporal information. The interventionist approach, however, needs to explain how temporal information fits in with intervention-based judgments. I will discuss this issue at length in the next chapter. For now, I will only specify that if my earlier suggestion of adding a temporal component along with the interventionist difference-making criterion is right, either as a claim about metaphysics, or as an additional platitude about our everyday concept of causation, then interventionism would have a wider area of applicability at least as far as psychological issues are concerned.

I would now like to move on to a different angle of looking at causal inference, interventionism, and the functional account with the aim of emphasizing another problem with focusing on an interventionist definition of the concept of causation. Reiss (2012a) presents a theory of the meaning of causal claims based on their inferential relations with evidence for causal claims. While I will not asses this view here, I would like to point to an interesting critique he makes against what seems to be commonplace in the literature on causation: ‘all standard accounts are “verificationist” theories (or developments of verificationist theories) in that they take conditions under which a causal claim can be tested for its truth or falsity to provide the meaning for the claim.’ (Reiss 2012a: 5) Reiss uses Woodward’s account as an example, where the meaning of ‘X causes Y’ is given by the fact that intervening to change the value of X would yield into a corresponding change in the value of Y. One problem with this, Reiss points out, coincides with one of the original problems that verificationism had to face: quite often, there are several ways of verifying the truth of a causal claim. In the case of the interventionist account of causation, while Woodward contends that manipulability is not the only way of making sense of causal claims, he does not have an explanation why other methods of obtaining evidence for causal claims are also working.

Reiss's criticism highlights a more general requirement that accounts of the meaning of causal claims should meet: they should be able to explain how causal inference can make use of several ways of obtaining evidence for causal claims. The standard accounts of causation, as he deems them, emphasize only one or a couple of features at most, but have trouble in explaining why causal relations may be inferred through different criteria. While this discussion is carried on the grounds of causal inference, the link to Woodward's consideration of a functional account of causation is obvious. It is in line with Woodward's (2014) considerations that an account of causation should provide one with an efficient means of inferring causally, or an explanation of why both intervention-based and non-intervention-based causal inference works.

One thing that is worth looking into are the examples of standard accounts that Reiss is using: the probability, counterfactual, process, and even agency accounts are supposed to be answering the metaphysical problem of causation; they also aim for a reductive analysis. One might ask how Woodward's account comes into the picture here. After all, in *Making Things Happen* he takes the effort to argue that, while being non-reductive and not sharing the same metaphysical goals, his account is preferable to both the counterfactual and the agency account. I believe that there are reasons to hold that unlike verificationism in the debate on the philosophy of language, the 'verificationist' feature of accounts of causation may rest on the idea that there should be some correspondence between what is taken to be the more fundamental feature of the world in terms of which causation can be analyzed, and the way in which causal relations come to be known. If Woodward is after a functional account of causation, there is no reason to limit the meaning of causal claims to a manipulability test. To be precise here, Woodward acknowledges that there are several kinds of evidence that support causal claims, but he proposes that they should be understood from the framework of manipulability. As Reiss puts it, ‘Woodward (...) merely claims that “reliable causal inference on the basis of nonexperimental evidence is possible”, tells us how to understand a claim inferred from non-experimental evidence [through hypothetical interventions – my note] but he owes us an explanation for why this should be so.’ (Reiss 2012a: 5)

A plausible explanation why Woodward’s account turns out to share this assumption with the standard accounts of causation is that his view on manipulability inherits a metaphysical assumption from previous accounts of causation that Woodward compares to his theory, and is supposed to connect the features of causal relations as they are out in the world with the human capacity of inferring causally. The fact that Woodward’s theory is vulnerable to the same kind of criticism is indicative of the fact that Woodward’s account does endorse some metaphysical assumptions, some of which have been pointed out in the previous chapters, and some other ones, which I will be discussing in the next section.

For the points that I will be making, an important conclusion to draw from Reiss’s criticism is that there are several ways of inferring causally and an account concerned with the meaning of causal claims in general should explain this fact. Moving the discussion in the context of causal reasoning and causal learning, the issue with interventionism is that, even though it acknowledges that there are several ways of identifying a relation as causal, it does not really say why it is so, and how that relates to the more general interventionist definition of causation.

As in the previous discussion about causal asymmetries a question that may arise is whether the interventionist account, as a functional approach to causation should deal with this issue. After all, as mentioned earlier, it is not required for it to work everywhere. There are two things that I would like to point out here, and the way in which they overlap might turn to be decisive with respect to the scope of a functional account of causation. The first is that once one subscribes to a functional account, i.e. an account that emphasizes the usefulness of causal claims, then the question is what concept/definition/approach works when talking about causation. In this sense, the scope is wider than an analysis of causation in terms of a more fundamental concept, since there are several ways of talking about causal claims, or of providing evidence for their truth. Secondly, since a functional account of causation has methodological rather than metaphysical aims, it can restrict its area of applicability to a set of relevant issues. In the case of Woodward's theory, the area of applicability seems to extend as far as the connection between causal claims and claims about manipulation and control go. Given the broader perspective on causation that a functional account aims to provide, the question is whether causation as manipulability should provide an explanation for other ways of obtaining evidence for or of inferring causal claims. Rather than giving a precise answer to this question, I would like to point out that supplementing the manipulability account of causation such that the asymmetry problem is answered would also broaden its area of applicability. Positing a link between manipulability, causal direction and time, either in the stronger or the weaker version can deal with a set of problems. Examples of such problems are:

a) Explaining why causal relations are asymmetric, or how people come to think of causal relations as asymmetric.

b) Accounting for other types of evidence for causal claims. I will particularly be concerned with the role of temporal cues in causal learning.

c) Dealing with cases of backtracking counterfactuals present in everyday as well as in scientific contexts.

To sum up my stance, I am not expressly claiming that the manipulability account should deal with these problems and limitations, but that having a broader area of applicability would be very much in the spirit of causal functionalism. I will now make a few observations on the constraints built into the manipulability account.

3.4 Metaphysical viewpoint constraints

Up until now, I have been pointing out some limitations of the manipulability account, especially in psychological context. In this section, I will explain them in the broader context described in the earlier sections. It will be the task of Chapter 4 to provide examples and draw further conclusions. In section 3.2, I have been pointing out that among the features of the Canberra plan there is the idea that everyday uses of causal claims should act as constraints on philosophical approaches to causation. A question to ask here is whether such constraints apply to Woodward's interventionist account. In the same section, I have also been pointing to the normative dimension of functionalism about causality: the criteria for intervention-based causal inference capture the way in which people ought to infer causally. The relationship seems to be the reverse of this aspect of the Canberra plan. In this section, I will make the claim that there are some different constraints that apply to Woodward's interventionist concept of causation, and that these constraints are more or less linked to the metaphysical assumptions endorsed at various parts of his version of interventionism. Before moving on, I should mention that my aim in this section is to investigate these constraints and explain their link to metaphysics. My endeavour is not in any way meant to give a metaphysical reading to manipulability, but rather, is tied to my earlier claim that, even when discussing more practical or methodological aspects of causality, there are some metaphysical assumptions and problems in the background.

A first constraint is the emphasis on intervention when defining causation. As pointed out above, from the perspective of a functional account of causation, there are several ways in which one can define causation, or provide evidence for causal claims. While manipulability is applicable to certain instances, a broader picture would also explain why other concepts are successful as well. In the upcoming chapter I will take a look at the various ways of inferring causal relations in psychological context. While Woodward (2011b) is mentioning an integration of a difference-making and geometrical-mechanical concept of causation, the relation between the two is not fully specified. As pointed out by Psillos (2004), at least as far as the debate between manipulability and mechanistic approaches to causation is concerned, the claim that one concept is better than the other still seems to be at stake. While my interest in the mechanistic approach to causation goes only as far as it intersects with issues concerning manipulability, I will specify the ways in which it is better at explaining certain experimental results with respect to causal reasoning and causal learning. I would, thus, claim that while a reductive analysis of causation in terms of manipulability or some other concept is no longer at stake, the debate between recent approaches to causation seems to have inherited some features of the debates between rival metaphysical accounts of causation. In functional context, however, the issue is not so much replying to counterexamples, but accounting for common practices in which causal claims are involved.

Another constraint on the functional aspect of manipulability is imposed by the commitment to realism. Woodward's approach to intervention, unlike earlier attempts to define causation as agency, is meant to be agent-independent. While Woodward does not explicitly discuss the relation between realism and objectivity, from the arguments discussed in chapter 1 it seems reasonable to assume that in Woodward's view the objectivity of causal relations is explained through their relation to features of the world and their independence from the agent's viewpoint. As I explained earlier, this pertains to using counterfactuals to spell out the conditions under which two or more variables are causally connected. How does this characteristic constrain the potential uses of the manipulability concept of causation? In the psychological, as well as in the philosophical literature there is work linking Woodward's concept of intervention to cognitive development data on causal learning. Woodward's concept of intervention is already complex enough to raise some worries about the extent to which it could match reasoning patterns in young children. However, the other part of the story is that Woodward's concept of intervention involves counterfactuals. A further task for the next chapter will be to look into data about counterfactual reasoning and see if a developmental account of causal learning in line with Woodward's considerations on interventionist counterfactuals can be built. As mentioned earlier, there is also a more pragmatic side to this constraint, and that is that it may be useful to think of causation as an objective relation. The problem is whether this aspect is reflected in the uses of causal claims in psychological context.

The final constraint that I am going to examine is the non-backtracking interpretation of counterfactuals. Even though I discussed the issue in philosophical context in the previous chapter, there is also the possibility of transferring it into the realm of causal understanding. The support of nonbacktracking counterfactuals in everyday contexts can be found in Woodward (2007). Once a claim about the psychology of causal inference and explanation is made, however, a look into the empirical data is necessary. As I will point out in the next chapter, there is experimental data showing that people sometimes backtrack when making inferences about what the cause of a given variable might be. From a metaphysical perspective, this constraint could be understood as an effort of making sense of causal relations as asymmetric (through having an asymmetric interpretation of counterfactuals). There is also a formal perspective inherent to Woodward's approach to causation through DAGs and his definition of intervention. I believe that a functional perspective should not exclude backtracking cases as long as they are informative and lead to useful causal claims or explanations. Once again, Woodward's approach does not have an account of why different ways of looking at counterfactuals may yield into useful results.

I believe that, at least partly, some of these constraints can be overcome by connecting manipulability to temporal direction. Once again, I emphasize that I do not take this move to be necessary for Woodward's account to hold as a very specific theory of causal judgments linked to manipulability and control. I do, however, think that linking it to a temporal component may broaden its area of applicability. In the next chapter I will illustrate how the previously presented constraints come into place and how a temporal component may help overcome some of them.

Chapter 4: Manipulability in psychological context

In the previous chapter I have singled out some features of the manipulability theory that may act as constraints when trying to apply the interventionist concept of causation to different domains of inquiry employing causal claims. In this chapter I will illustrate how these constraints work when applied to psychological work on causal reasoning and causal learning. This chapter will have three main parts, corresponding to the constraints I have identified. The first two, in particular, will focus on conceptual development data. The third will consider some experimental data on counterfactual reasoning as well as an inquiry into a model of explanation in the social sciences. The aim of this chapter will be to illustrate how these constraints operate and how some of them could be overcome either by connecting manipulability to temporal direction or by acknowledging a connection to an agency concept of causation.

To recapitulate, the constraints that I have identified are:

1. The focus on manipulability. Defining causation strictly in terms of manipulability might leave out some aspects of causality (or, as Psillos 2004 puts it, ‘causation has excess content over invariance-under-interventions’). Although from a functional perspective, the manipulability account is not required to work everywhere, given the emphasis on the usefulness of causal claims, it still seems to owe us an explanation as to why other ways of inferring causal relations are successful.

2. The objectivist definition of intervention, namely the use of counterfactuals. While not a constraint stricto sensu, in psychological, especially developmental context, it may be difficult to distinguish this specific concept of intervention from a more subjective one (i.e. one that does not involve counterfactuals).

3. The non-backtracking interpretation of counterfactuals. While I have pointed out that in the broader context of the problem of causation this interpretation might connect to the asymmetry issue, from a functional point of view one may still raise the question why some uses of backtracking counterfactuals in causal reasoning yield informative results.

4.1 Causality in philosophy and psychology: general remarks

I will start by making a few general observations on causality in philosophy and psychology. I will first try to explain how philosophical approaches, the manipulability theory in particular, relate to experimental work on causal reasoning and causal learning. I will then proceed to spell out some distinctions concerning causality in psychological context that will be useful for the subsequent discussion.

4.1.1 Philosophical and psychological perspectives on causation

As recent literature on causation shows, philosophical approaches to the problem of causation and experimental inquiries into causal reasoning and causal learning can interact, resulting in new perspectives on causation and causal understanding. I will sketch an overview of the connections between these two fields in their investigations of causation and causal concepts. My analysis will mainly focus on Woodward's treatment of causation and its applicability to experimental work into causal reasoning and causal learning. Relying on the previous considerations on the applicability area for manipulability, I will proceed to raising some questions concerning some of the claims from Woodward's work as well as from the literature on causal reasoning.

A first thing to clarify is the scope and aims of each one of the two perspectives. Looking into what the philosophy of causation has to offer, the most general issue is how causation should be defined. The main question here concerns the more fundamental concept in terms of which causation should be analysed. While various answers have been given, all of them have been struggling with counterexamples or with the inability to handle various cases. Among other things, this has motivated philosophers to opt for approaches that do not aim at reducing causation to a more fundamental concept, but, rather, at presenting a framework accounting for the way in which causal claims work in both everyday, as well as scientific contexts, and at explaining how causal structures come to be known.

Due to his non-reductive definition of intervention along with the emphasis on methodology, Woodward's theory fits in with the latter project. Furthermore, as discussed in the previous chapter, Woodward (2014) defines his project as functional, in the sense that it is oriented towards the usefulness of causal claims for various cognitive purposes. There are some things that Woodward's approach shares with the broader, metaphysical project, however. One of them is the concern about ontology: even if an account focusing on the workings of causal claims in different contexts does not need to take a stance on metaphysical issues, Woodward nevertheless subscribes to a realist view. On his account, the truth of intervention-based causal claims is grounded in some features of causal relations as they exist in the world. Woodward employs counterfactuals when defining his concept of intervention in order to support a concept of causation that is objective, independent of the subject's goals and capacities. For this reason, I take Woodward's account to bring a broader perspective on causation, going beyond a normative account of causal inference and its applications to an experimental setting. To put it another way, the interventionist account of causation explains how causal claims come to be useful for certain purposes and goals, but it also provides a realist answer to the question as to why causal claims based on interventions come to be useful. On the other hand, it is also worth pointing to the structural similarities between Woodward's approach to causation and the Bayes nets approaches to causation as long as one keeps in mind that the scope of Woodward's theory goes beyond offering a formal model of how causal structures can be discovered.

On the psychology side, there lies a different set of issues. Some of them concern developmental investigations into the concept of causation. Recent work on causal learning aims at providing a causal Bayes net model for causal learning in young children. There are two main such approaches: Bayesian methods and constraint-based methods. Bayesian methods, as in Griffiths & Tenenbaum (2009), use Bayes' rule on prior beliefs about the probabilities of different causal structures. The beliefs are updated according to evidence based on observation or interventional data and to how probable that data is taken to be given the structures. Constraint-based models (Sprites, Glymour & Scheines 1993, Gopnik et al. 2004) use statistical data to establish conditional dependencies and independencies. Given the conditional probability relations, certain structures are ruled out. While I will not go into details about these models, it is worth pointing out that Gopnik and her collaborators argue for the constraint-based model as a model for human learning. An important thing to stress is that, on both models, interventions can be part of the evidence data.

Since, as pointed out above, Woodward's approach is consistent with Bayesian net approaches to causation[56], particularly, with the constraint-based model used by Gopnik et al., the hypothesis that children learn causal structures in conformity with a principle that closely matches Woodward's definition of intervention is worth looking into. However, there is experimental evidence that Bayesian reasoning in causal contexts has its limits, and there is also experimental work linking causal reasoning to other approaches, or concepts related to causation, such as mechanisms, or temporal cues. These issues are not limited to the conceptual development domain, but apply to a wider area of psychological investigations. A first question that I will be addressing, taking the empirical evidence into consideration, is to

[56] Woodward particularly refers to the model used by Gopnik et al., but, as specified earlier, interventions work in the Bayesian model as well.

what extent would a model of causal reasoning analogous to Woodward's approach to causation as intervention be applicable to developmental studies into causal reasoning. The second question that I will be concerned with is how can success on causal reasoning tasks based on intervening be associated with Woodward's specific approach to intervention, given the complex conditions that Woodward sets out for a variable to count as an intervention.

Before proceeding to investigating these issues, I will dispel a worry that philosophers may raise. One may ask why the way in which children learn about causal relations should be relevant to the philosophical inquiries into the metaphysics or epistemology of causal relations. One answer is that, ultimately, I will be addressing a philosophy of science question here, namely whether the interventionist concept of causation provides a good model for the development of causal learning, and thus whether it could help developmental psychologists in shaping experiments and hypotheses concerning causal learning. A closely related way of putting it would be to inquire whether the interventionist concept of causation provides an accurate description of the psychologists' assumptions about what causal relations amount to in experiments on causal reasoning and causal learning. This investigation could also be helpful in relation to the fundamentally pluralist idea that there are different concepts of causation fitting certain domains, although I will not undertake such an investigation here. From such a perspective, the question would be how well interventionism fits in with investigations into causal reasoning and whether it is causation as intervention or another notion of causation that may do a better job in this context.

As to how are the philosophical and psychological issues concerning causation related, from the viewpoint of Woodward's account the answer is fairly simple: 'if we have a normative theory that tells us that we ought to reason about causal relations in certain ways (...) and if we find people in fact reasoning in some good approximation to what is recommended, then these facts can form part of a potential explanation of why (and to what extent) people are successful causal reasoners.' (Woodward 2012: 962-963) As specified above, Woodward's normative account is rooted in an objective concept of causation. Finally, it should also be mentioned that the assumption that research into causal learning in young children is relevant for the adult concept of causation is endorsed by Woodward in his work on the interventionist concept of causation in psychological context (see Woodward 2007; 2011b). Furthermore, in recent debates around causation the problem of the origin of causal concepts is sometimes relevant. For instance, in Gijsbers & de Bruin (2014) Woodward's interventionist theory is criticized on the basis of a 'genesis problem': 'although Woodward can hold that his theory captures the meaning of causation, the theory nevertheless makes it highly mysterious how we could ever acquire such a concept and start gathering causal knowledge.' (Gijsbers & de Bruin 2014: 1776). The authors' solution, which involves the claim that the interventionist concept of causation is derived from the agency one, includes an important psychological component, namely the development of causal concepts and some empirical evidence in favour of the agency concept. While I will be coming back to this argument in section 4.3, the point I wish to make here is that the origin of causal concepts along with empirical evidence plays an important part in current debates around causation and manipulability.

4.1.2 Causality in psychological perspective: two distinctions

I will now present two distinctions that Woodward uses in his work and that will be relevant for the points I will be making in the sections to follow. One distinction is that between different stages of causal understanding presented in the context of applying the interventionist concept of causation to research into causal understanding in young children. According to Woodward 2007b, these stages are:

(1) egocentric, where the agent grasps causal relationships, but does not recognize the presence of the same relationship in the absence of the agent's acting;

(2) the agent causal viewpoint, where the agent grasps the same relationship when other agents are involved;

(3) the fully causal viewpoint where the agent grasps the same relationship in the presence of other agents as well as when no agent is involved.

In a further article, Woodward (2012) explains that this distinction provides new ways for psychologists to design experiments. A study that Woodward co-authors (Bonawitz et al. 2010) discusses experiments on how toddlers and pre-schoolers are able to bring about an event (making a toy airplane move) after watching the event brought about in one of two conditions. In the agent condition a hand moved the block towards the plane, whereas in the ghost condition the block moved by itself. Upon being asked to make the plane move, pre-schoolers succeeded in both conditions, whereas toddlers failed in the ghost condition. In Woodward (2012) this is taken to show that pre-schoolers developed a fully causal viewpoint, whereas toddlers were at the agent causal viewpoint, even though they showed understanding of the causal connection in both cases.

This distinction provides a useful way of fitting the philosophical insights between causation and intervention to the context of developmental investigations into causal cognition. While Woodward emphasizes the importance of the interventionist concept of causation with relation to early causal learning, it would be difficult to attribute young children the full blown apparatus behind interventionism. Thus, the fully objective understanding of causation and interventions is only present later on, when the fully causal viewpoint comes in.

Another distinction that will be of use in the forthcoming investigation is that between difference-making and geometrical-mechanical concepts of causation. While I have discussed this distinction into more detail in the previous chapter, I will make use of some aspects of the debate concerning a counterfactual versus a mechanistic model of scientific explanation, specifically, how they relate to cognitive development data.

4.2 The interventionist concept of causation, temporal cues and mechanism information

I will now look into some psychological work concerning the development of causal concepts and see how Woodward's concept of causation fits in with the experimental data. I will first present a set of experiments and conclusions that seem to support Woodward's definition of intervention and then point to further experimental evidence that may render it problematic. Subsequently, I will look into the relation between interventions, mechanism information, and temporal cues. In the light of the experimental data, I will explain how the constraint from defining causation as manipulability comes in. Finally, I will point out how connecting manipulability to temporal direction may be of help in providing an account for successful causal inference based on temporal cues.

4.2.1 Cognitive development and the conditional intervention principle

I will first investigate how work based on probabilistic models by Gopnik, Schulz, or Glymour, among others, is compatible with the interventionist concept of causation. I will first discuss an experiment by Gopnik, Sobel, Schulz, and Glymour (2001) where, through observations, children are shown to be using probabilistic evidence for identifying a causal connection. In this experiment, 3 and 4-year olds were presented with a device called a 'blicket detector'. They were further told that some objects are blickets whereas others are not and that 'the blickets make the machine go' (the machine was activated by the experimenter through a hidden device and it made a sound whenever a 'blicket' was put onto the machine). The children were then presented with a scenario where two objects A and B were simultaneously put on the detector and the sound would go off. Afterwards, they were presented with A and B separately; A would activate the machine, while B would not. The majority of the children would say that A is a blicket, whereas B is not a blicket. This shows that while presented with two possible causes, and with the evidence about conditional probabilities (the probability of the machine's activation is dependent on A, or, as the authors explain, A screens off B), the children would single out the right cause.

A second experiment involved children being presented with two objects, A and B, separately placed on the machine. A would activate the machine three times in a row, out of three, while B would activate the machine only two times out of three. Asked which objects are blickets, 97% of the children said A is a blicket and 85% of the children said that B is a blicket as well. This result further shows that children connect causal claims to increase in probability. While B activates the machine only in 66% of the cases, it is still taken to make a difference towards the activation of the machine.

These experiments are consistent with Gopnik's claim that children learn according to probabilistic models. In the 'blicket detector' scenario, children show they 'can and do infer new causal relations from information about dependent and independent probability.' (Gopnik et al 2001: 628). Turning to how these findings correlate with a more general account of causation, one could notice the similarity with probabilistic accounts (causal claims are based on information about probabilities) and with a Bayesian picture of evidence (upon observing certain instances, children infer specific causal relations; other types of experiments show that children also update their past causal judgments when presented with new evidence). An issue with this kind of experiments is that while they rely on judgments about probabilities in order to establish a causal connection, the children are not involved in any interventions over the system. Thus, the causal claims they make are exclusively based on observation.

In another series of experiments by Schulz, Gopnik, and Glymour (2007), children are shown two gears A and B and a switch S. When the switch is in the 'on' position both gears rotate. As there are three possible scenarios considered, common cause (A<-S->B), and causal chain (S->A->B, or S->B->A), children were presented with evidence that would support one of these scenarios. For instance, in the S->A->B case, the experimenter turned the switch off, then removed A and, upon turning the switch back on, children could see that B no longer rotated. Upon removing B (and replacing A), A would still rotate when turning on the switch. After being presented with this kind of evidence, the children were able to choose the correct causal structure, depictured in accordance with the possible scenarios. In another experiment, children were able to play with the gears themselves and establish the causal structure. They were successful especially when working in pairs. In yet another experiment, children were presented with the causal structure and then they were able to predict what would happen on certain interventions.

The conclusions that the authors draw from this set of experiments are that children are able to identify causal structures through interventions by the experimenter, as well as by intervening themselves on the variables in the system. Furthermore, the authors consider that the results could be subsumed under what they call the conditional intervention principle, which requires one to:

1. Hold all the other variables in the system fixed.
2. Have an intervention on X that
3. will change the probability distribution of Y
4. but not influence Y other than through X
5. and not change the fixed values of the other variables in the graph. (reconstruction after Schulz et al. 2007: 323)

The principle closely matches Woodward's concept of intervention, rendering plausible the supposition that the authors use criteria similar to the ones in Woodward (2003) in order to attribute success on causal reasoning tasks to young children. Another conclusion to be drawn from these findings is that children learn about causal structures in accordance with the requirements set by Woodward's theory, as he makes it explicit that 'people learn and reason in accord with the normative requirements of the interventionist account' (Woodward 2007:

28) Thus, on a first glance, Woodward's theory's seems to provide a useful framework for developmental investigations into causal reasoning. Furthermore, it also seems to set up a principle that governs causal inference in young children.

Before going on to examining experimental evidence that may suggest otherwise, I should point to a few problems that the latter set of experiments may exhibit. One issue, pointed out by authors of subsequent studies is that although there is an on/off switch, in this task children are basically working with a two-variable system. In order to do justice to the interventionist claim about causation, a system of at least three variables would be necessary.[57]

A more general issue that I wish to raise at this point, but that I will be discussing into more detail later on, concerns how the conditional intervention principle should be interpreted. There are instances, such as the ones at use in the experiments described above, where it successfully applies to children's judgments about causal relations. However, as I will show shortly, there are instances where its applicability may be undermined either by a failure to make consistent judgments about causal structures and interventions, or by a preference for other ways of inferring causal structures. I will explore these problems in next subsection.

4.2.2 Interventions, temporal cues and mechanisms

An important worry about the conclusions of the Schulz et al. (2007) investigation is raised from further research into causal learning based on intervention. A study by Frosch, McCormack, Lagnado and Burns (2012) shows that children between 4 and 8 years old presented with a causal structure (either inferred by the children, or given by the experimenter) give inconsistent answers about the causal structures and the effects of

[57] As pointed out previously, Woodward's concept of intervention requires that the arrows between the variable intervened upon and its parent be broken. In order to see if one is reasoning according to this claim, the system should include, besides the putative cause-effect pair, the variable that has a causal influence on the putative cause.

interventions. In a set of experiments, children are presented with three variables represented by three objects of different shapes and colours on a box. Their interactions correspond to one of the common cause or the two causal chain scenarios. In the first two experiments, the children had to infer the causal structure from temporal cues (for instance, in the common cause scenario A starts moving, and a few seconds later B and C start moving; in the ABC causal chain A starts moving, followed by B, followed by C). After inferring the causal structure, children are asked if B would still move if C were stopped from moving, and the other way around. After receiving inconsistent answers after Experiment 1, in Experiment 2 only the older group of children had to infer the causal structure and had to answer future questions, rather than counterfactual ones about interventions. In Experiment 3 children were told the causal structure, and then asked questions about interventions, but the overall performance was still poor.

Without going into the full details of the results (there were some discrepancies in the answers to the common cause and causal chain scenarios), what is relevant to point out is that the children's intervention judgments did not consistently correlate with the causal ones. This means that children may not yet see the connection, or that they may think of causation in terms different from intervention. These results go against what the authors call the strong interventionist claim from Schulz et al (2007) that 'a causal relation is defined (...) in terms of the real and counterfactual interventions it supports' (69), and that 'when children infer that a relationship is causal, they commit to the idea that certain patterns of interventions and outcomes will hold' (70). An interpretation suggested by Frosch et al. is that 'children may find it difficult to give coherent answers about the effects of intervening on components of the system without knowing anything about the underlying mechanisms that connect the components.'

This may be a good point to emphasize that there are various ways in which children could infer causal structures. The ‘blicket detector’ scenario presented above seems to rely on probability information, the experiments involving gears rely on interventions, whereas the causal scenarios in the Frosch et al. article rely on temporal cues. As I will be discussing in the next subsection, there is also experimental evidence that children can infer causal structures through information about mechanisms. Coming back to the issue of the conditional intervention principle stated above, the failure to link causal judgments to intervention judgments seems to point out that the principle may be limited to structures simpler than a three-variable system. However, children are able to infer those causal structures by different means. Thus, a further interesting issue would be to make sense of how all of these criteria for causal inference fit in conceptual development. For philosophers, a thing to note is that concepts that have been used in providing various definitions of causation can serve as criteria for causal inference. While investigations into causal reasoning will not provide one with information as to what concept is more fundamental from a metaphysical perspective, it may provide information with respect to what is understood in terms of what.

I will now present some findings from a study by McCormack, Frosch, Patrick, and Lagnado (2014) discussing temporal cues, probability information and interventions in the context of causal reasoning. The experiments involved inferring causal structures from a device similar to the one from the Frosch et al. (2013). Participants were groups of children from 5 to 6 years old, from 6 to 7 years old, from 7 to 9 years old and adults. In Experiment 1, participants had to infer a causal structure after being presented with incongruent evidence from temporal cues and statistical information. In Experiment 2, the conflicting evidence came from temporal cues and intervention information (in this case, intervention pertained to activating each one of the variables in turn and seeing what other variables would simultaneously activate). In Experiment 3, temporal information was not available, and participants had to infer the causal structures from the statistical data about intervening to disable one variable and activating the other ones in turn. Without going into too much detail about the procedure, the results show that children preferentially employed the temporal cues for inferring the causal structures in the incongruent cases. In the last experiment, where no temporal cues were available, only adults and the older group of children performed above chance. The authors provide two explanations of these findings. One is that statistical information is more difficult to process, and thus children mostly relied on temporal cues. The other explanation is that there may be an inherent bias in children as well as in adults for using temporal information over statistical information when inferring causal structures. The authors connect this to two studies by Shulz (1982) and Ahn et al. (1995) that argue that children and adults place more weight on mechanism related information than on statistical information, which suggests that the geometrical-mechanical model of thinking about causation may be more fundamental than the difference-making one. The authors point to a link between the geometrical-mechanical account of causation and temporal information. If this is the case, then the preference for temporal cues also points to the geometrical-mechanical notion as being more fundamental.

Before going deeper into the geometrical-mechanical versus interventionist accounts of causation issue, there are some further things to mention. One of them concerns the connection between causal judgments and temporal judgments that the empirical results point to. Once again, another look at the metaphysics of causation shows associations between the nature of causality and temporal direction. As presented in chapter 2, there are several attempts to explain both the direction of time and the direction of causation through another concept (e.g. counterfactual dependence or the direction of entropy). The geometrical-mechanical theories of causation also may be interpreted as including a temporal component. Finally, in chapter 2, I have subscribed to the solution of explaining the asymmetry of causal relations through the asymmetry of time. The question is whether temporal information connects to the interventionist take on causation. Making use of the distinction between the geometrical-mechanical and difference-making notions of causation, the issue seems to be that temporal order is not a criterion for difference-making. Seeing A being followed by B does not have any bearing on whether A's occurrence makes a difference to B's occurrence. But why would temporal succession be an indicator of a causal relation? My suggestion is that the use of temporal cues to infer that A causes B connects to the idea that causation has a direction, and that direction follows the arrow of time. While this relation is not reflected by exclusively observation-based statistical information, it can fit in with an interventionist model: one can intervene to alter the future through altering the present, as well as one can intervene to alter effects through altering their causes, but not the other way around. Without expanding on this issue[58] here, a point to take from philosophical investigations on the asymmetry of causation is that intervention (or at least Woodward's concept, for my purposes here[59]) is making use of this asymmetry rather than grounding it. A similar point is made in psychological context by Lagnado and Sloman (2004). In a set of experiments that involve inferring causal chain or common cause structures, people proved to be more successful at both intervention as well as observation based inferences when they were intervening on/observing systems where there was a temporal delay between the cause and effect variables. In the general discussion, the authors emphasize the advantage of intervention over observation in terms of ruling out confounders. They also point out that in everyday contexts interventions are prior to their effects. It seems, thus, that in this setting intervention and temporal information come in as a package of features accounting for the everyday concept of

[58] There is also evidence that the relation between causal judgments and temporal judgments goes both ways, with judgments about causal structure determining judgments about causal order, see Bechlivanidis & Lagando (2013).

[59] This comes as straightforward since on Woodward's theory intervention is a causal concept. As I have pointed out in chapter 2, the issue could be generalized over other approaches that seek to ground the asymmetry of causation in manipulability.

causation. These findings are consistent with my claim about the connection between temporal order and the direction of causation and between intervention and difference-making information. In the light of the previously presented data, where temporal and intervention information are conflicting, the fact that people are biased towards temporal cues when inferring causally seems to suggest that the understanding of the direction of time is more fundamental than the understanding of the direction of intervention.[60]

Coming back to the issue raised at the end of the previous subsection, if the temporal and mechanism judgments are not only present in causal learning, but are likely to be more fundamental in the development of causal thinking, where does that leave the conditional intervention principle? As McCormack et al. point out, the causal Bayesian framework of the studies of Gopnik and her collaborators is a normative one, so children do not need to always make judgments about causal structures based on statistical information provided by interventions. One of the conclusions of the McCormack et al. study is that it is not easy for children to make such judgments. Thus, coming back to the question stated in the beginning, there seem to be some doubts over how well the interventionist way of inferring causal structures may fit in with developmental studies. While there is evidence that older children and adults successfully make use of interventions, the capacity to make causal judgments based on statistical information about interventions seems to emerge at a later point in development. But, from a philosophical viewpoint, why is it important which one of them is developmentally more fundamental? To answer this question, I am going to draw some relevant points from the debate between Woodward and Waskan on causal explanation, mechanism and cognitive development data.

[60] This could also be part of the explanation of the causal vs. diagnostic reasoning asymmetry discussed in chapter 3.

4.2.3 The interventionist versus the mechanical necessitation account

One thing to mention is that the debate concerns the counterfactual and mechanistic models of causal explanation. I will not focus on explanation here, however, but rather on the assumptions that the discussed views hold about causality and about the importance of developmental data on the origin of the concept of causation. Waskan (2011) points to some experiments by Schlottman (1999) on causal perception in 5-year olds, 7-year olds, 10-year olds, and adults. On one experiment, they were presented with two balls being dropped into separate holes of a box. The second ball was dropped 3 seconds after the first ball. After the second ball was dropped a bell inside the box rang. Upon being asked what caused the bell to ring, all participants showed a strong preference for the second ball as causing the bell to ring. Afterwards, the participants were familiarized with two mechanisms, a see-saw device that would make the bell ring immediately after being touched by the ball, and a ramp that would require the ball to roll for a few seconds before touching the bell and making it ring. After placing one device at a time under one of the holes in the box, all the participants were able to correctly predict how long it would take for the ball to make the bell rang. Finally, the participants were presented with one of the devices being placed inside the box (though they couldn't see under which hole) and with the two balls being dropped with a 3 seconds delay between the first and the second ball. In the case of the ramp device adults correctly judged the first ball to cause the bell to ring, 10-year olds performed pretty much like the adults, for 7-year olds performance was intermediate, whereas the 5-year olds answered that the second ball caused the bell to ring. The conclusion is that at a younger age causal perception (given by temporal information) comes before mechanism information. Later on, this relation is reversed.

Waskan's interpretation of this data is that 'causal perception is triggered by certain forms of temporal contiguity information (e.g., involving spatial, sonic, and presumably other properties) and the application of our concept of causation as occurs in cases of causal belief is triggered either by causal perception (preferentially so in young children) or by either superficial or deep justificatory information (preferentially so in adults).' (Waskan 2011: 399) Waskan then proceeds to argue for a concept of causation as mechanical necessitation. Based on the empirical data, he suggests that 'early in childhood it might amount to the idea that there is a class of circumstances in which the prior occurrence of X in some underlying or mediating system of (at least) spatiotemporally arranged parts is connected to the subsequent occurrence of Y through a chain of expected behaviors (…) (e.g., regarding impenetrability, collisions, support, etc.).' (Waskan 2011: 402)

One of the points made by Woodward (2011b) in his reply is that the adult notion of causation cannot be fully derived from causal perception and information on mechanisms, and that it needs to integrate considerations about difference-making. One thing worth pointing out here is that Woodward is both defending the distinction between geometrical-mechanical accounts and difference-making accounts, and emphasizing how the adult notion of causation should contain both elements. According to Woodward's critique, Waskan's view, along with his considerations on developmental data, seems to suggest that the full-blown adult concept of causation is based solely on the elements belonging to the geometrical-mechanical notion.

An interesting thing to note is that developmental studies into causal learning are important for both Woodward's and Waskan's views on causation and causal explanation. Bootstrapping from causal perception to the adult notion of causation, and the way in which geometrical-mechanical or difference-making considerations on causation are involved, is by and large an empirical question. Once again, the issue here seems to be generality: can the mechanical necessitation account view be accounted for in terms of a counterfactual view, or the other way around, or at least which one has a broader area of application? I am now going to draw some conclusions from the interactions between the interventionist version of difference-making and developmental data on causal learning.

This would be the right point to assess how the constraint relating to the definition of causation as manipulability applies to the issues discussed in the last two subsections. As far as the work linking Woodward's concept of intervention to developmental data on causal learning goes, the worry is that young children may not successfully apply interventions on a three-variable system. Nevertheless, as the empirical data presented above shows, they are able to infer causal relations on the basis of temporal cues. More data on causal perception shows how younger children infer causal links on the basis of temporal order, while later on they do so on the basis of mechanism information. A very interesting question for developmental psychologists concerns how one reaches the adult concept of causation. An important thing to note from a philosophical perspective is that the adult concept of causation is associated with both intervention and mechanism information. Since Woodward relies to some extent on an analogy between the development of causal concepts and the structure of causal explanation, it is important to investigate to what extent his concept of causation fits in with developmental investigations. The difficulties of applying the interventionist concept of causation to developmental data are twofold. On the one hand, as specified earlier, the manipulability theory accounts only for those uses of causal concepts that involve actual or potential interventions. If one is to make sense of the various ways in which causal structures are inferred and connect them to the manipulability concept of causation, then an explanation of why they work is necessary. On the other hand, Woodward's version of manipulability brings in a set of complex features: thinking of causation in terms of DAGs, counterfactuals, arrow-breaking interventions, which would be difficult to attribute to early stages of causal understanding.

Looking at these issues through the distinction between causation as difference-making or as production, another thing to point out is that the geometrical-mechanical concept can make sense of temporal information. The complexity problem could be dealt with according to the specific way in which mechanisms along with their relation to causation are defined. Thus, in this particular respect, the geometrical-mechanical concept of causation proves to be more general. My suggestion is that incorporating temporal contiguity along with manipulability information could deal with the former issue and, in this respect, bring the manipulability concept of causation on a par with the geometrical-mechanical one. The idea that causes precede their effects seems to be developmentally more fundamental than information about interventions. However, as noted above, temporal information is not sufficient for a relation to be causal. Thus, a full blown concept of causation incorporates both temporal order and difference-making information in terms of interventions. The fact that temporal information comes in earlier than manipulability information is consistent with the earlier observation that the interventionist concept of causation makes use of the asymmetry feature (understood in terms of temporal direction) rather than explaining it. I would also like to say a few more things on the issue of simultaneous causation. Leaving aside the metaphysical considerations from chapter 2 and conceding that simultaneous causation may be possible does not change the way in which causal relations are ordinarily understood. If temporal order is taken to be indicative for causal order (and the experimental work quoted above shows that it is), then simultaneous causation may be a concept harder to grasp in everyday contexts. Once again, experimental evidence (see Lagnado and Sloman 2004 and McCormack et al. 2014 discussed above) shows that people find it harder to identify causal structures in the absence of temporal cues. I suggest that admitting the possibility of simultaneous causation should not influence our general way of understanding causal relations. Rather, it would imply that there could be some particularities about reasoning with causes that occur at the same time with their effects. This would also be a good point to note that integrating temporal and manipulability information would help account for inferring causal structures based on temporal cues, as well as for explaining the temporal component of mental simulation that differentiates cause to effect from effect to cause reasoning.[61]

Nevertheless, the complexity problem still stands. Before discussing it in more detail, I will use the next section to elaborate on one of its aspects, namely the link between causal reasoning, counterfactual reasoning and intervention.

4.3 Causal and counterfactual reasoning in developmental context

In this section and the next one I will be focusing on counterfactual reasoning. As specified earlier, my aim is to illustrate how the constraints identified in the previous chapter apply to Woodward's version of manipulability when discussed in psychological context. In the current section I will be addressing two questions. The first is whether current developmental investigations into counterfactual reasoning support a concept of intervention along the lines of Woodward's definition. The second question concerns how should the concept of causation associated to the experimental tasks (most importantly, its objectivity) be understood if children do not show a satisfactory understanding of counterfactuals. Concerning the relevance of such inquiry, I would like to point out that it is a continuation of the previous investigation of manipulability as a good model for experimental work on the development of causal concepts. While in the previous section I have investigated how intervention fares by comparison with other ways of inferring causally, in this section I will investigate whether interventions are connected to counterfactuals in cases where intervention-based causal

[61] As discussed in chapter 3.

reasoning is at work. Based on empirical data,[62] my answer will be negative. However, I will be pointing to a few ways in which someone supporting an interventionist and objective concept of causation may find a way around the issues concerning counterfactual reasoning.

4.3.1 A link between causal reasoning and counterfactual reasoning

In a 2009 article, McCormack, Butterfill, Hoerl, and Burns investigate, among other things, how the causal claims that children infer from the blicket detector paradigm correlate with counterfactual judgments. They classify the trials in backward and forward and generative and blocking. In the backward trials children are presented first with two objects, say A and B on the detector, then only with one, either A or B. In the forward trials, they start with either A or B on the detector and then A and B together. In the backward trials they first witness A and B activating the detector, and then only one object activating or failing to activate it whereas the forward trial goes the other way around. In the generative scenarios, children are presented with evidence that an object is not a blicket (say, B fails to activate the machine), while in the blocking scenarios they are presented with an object (A) activating the machine. Using these scenarios, the authors manage to obtain causal judgments from groups of 4-year old and 5 to 6 year-old children. In the last two experiments, after presenting children with such scenarios, they ask them counterfactual questions of the form 'A moment ago I put both these blocks on the machine together, and it went off. Do you think it would have gone off if I hadn't put this [gave color of A] one on?' (McCormack et al. 2009: 1570) in Experiment 3 or 'Do you think the machine would have gone off if I had only placed B on it?' (McCormack et al. 2009: 1571) in Experiment 4. Without going into the details of the study, the important conclusion for the problem discussed in this chapter is that the children's answers of correctly classifying the objects as blickets, and thus attributing them causal powers correlate with their answers to

[62] I should point out that Woodward also makes use of experimental data as support for his take on counterfactuals (see Woodward 2007). Since his claims concern the problem of backtracking in adult causal reasoning, I will be discussing this in section 4.4.

the counterfactual questions. The authors' point is that the blicket detector paradigm can account not only for the capacity of making causal inferences, but also for counterfactual reasoning in children.

Another interesting remark is that these causal claims could also be expressed in terms of conditional probabilities. However, as the authors point out, the probability relations are difficult to grasp even for adult participants in similar scenarios. Thus, understanding causal relations in term of counterfactuals, or at least connecting the two kinds of judgments may be a more accessible path than the one described by Gopnik and colleagues.

A thing that may spring to mind at this point would be that on this approach children are shown to have consistent judgments about causal structures and counterfactuals. However, unlike some of the cases described by Woodward, they do not involve direct intervention by the children, as the blicket-detector paradigm relies on children observing interventions made by the experimenter. The general claim that this data supports is that there is a connection between causal and counterfactual reasoning.[63] The way in which it could support an understanding of the concept of causation along the lines of Woodward's theory is that it seems to show that children do employ counterfactuals in causal contexts. However, it could also be consistent with any view that takes counterfactuals to be a useful way of spelling out difference-making information.

4.3.2 Counterfactual reasoning and conditional reasoning

I will now present some experimental data that may provide reasons for doubting that children reason counterfactually. Work by Perner and Rafetseder, among others, points to the conclusion that previous success attributed to children in tasks involving counterfactual reasoning can be attributed to conditional reasoning.

[63] Once again, I emphasize that I am focusing on developmental data here. Otherwise, there is support that counterfactuals play an important role adult causal reasoning (see Gerstenberg et al. 2015).

In a study by Rafetseder, Cristi-Vargas and Perner (2010) children of 4 to 6-years old are presented with stories on the basis of which they are required to answer future indicative (relating to conditional reasoning) and past subjunctive (relating to counterfactuals) questions. Discussing the scenario and the questions used in the second experiment would be sufficient for the purposes of this section. The children are presented with a story where a mother leaves sweets either on the top shelf (L1) or the bottom shelf (L2) of a drawer. Either a tall boy (C1), or a little girl (C2) can come and try to take the sweets on the following condition: the boy can pick up the sweets from the top shelf, but not from the bottom one, because he has hurt his leg before and cannot kneel to reach the shelf; the girl can only pick up the sweets from the bottom shelf since she is not tall enough to reach the top one. If either the boy or the girl picks up the sweets, they will take them to their room. The questions differ, according to the different possibilities of the scenario, the main forms being (in total, there are four such questions, as the number of relevant variables: L1/L2 and C1/C2):

Indicative future: what will happen to the sweets if the boy comes looking for them? (when they are on the top shelf)

Subjunctive past: where would the sweets be if the little girl, and not the tall boy, would come looking for them? (when the boy comes along and takes the sweets to his room)

The results show a certain discrepancy between indicative future, where most children got the answers right and past subjunctive, where performance dropped significantly. They did especially worse in the condition present in the question above, where children answered that the girl would take the sweets to her room, even if they were on the top shelf. As the authors put it, ‘to get it right, children have to consider where mother had actually put the sweets. In other words, they have to construct a possible world that is maximally similar to the actual world, one that takes account of where the sweets had actually been put.’ (Rafetseder et al 2010: 382) According to these results, children fail to do so, and only a small number manage to get the whole four counterfactual questions right. A further experiment shows that adults are successful in this task.[64]

The explanation that the authors provide, and which will be interesting given previous talk about a link between causal reasoning and counterfactual reasoning, is that before the age of 6 children are able to operate with conditionals, but not with counterfactuals. It is conditional reasoning that explains the success of children in previous experiments from which they seem to exhibit counterfactual reasoning.

In a further study, Rafetseder and Perner (2010) make further observations on counterfactual and conditional reasoning. One of their claims is that in order to to answer counterfactual questions correctly, children need (1) to create an alternative scenario to reality and (2) to integrate the information about what has happened with information about what could have happened. Finally, in a more recent study, Leahy and Perner (2014) further elaborate on the distinction between counterfactual reasoning and basic conditional reasoning from the point of view of the information taken into account: 'when faced with a counterfactual question about a story whose antecedent contradicts a nonpermanent feature of the story, CF [counterfactual] reasoners take into account both permanent and nonpermanent features of the story. BC [basic conditional] reasoners presented with the same question only take permanent features of the story into account.' (Leahy & Perner 2014: 803)

Having presented the experimental findings, it is time to address the way in which they relate to interventionism about causation. One obvious conclusion to draw about developmental data is that, since they cannot reason counterfactually, young children cannot

[64] It may be objected that the requirements in terms of similarity between possible worlds, as stated by Rafetseder et al., are too strong, even for adults. This would further entail that the talk of counterfactual reasoning would be useless when trying to see whether an approach such as Lewis's may be right. My interpretation of the passage is that the talk of possible worlds need not take the shape of possible worlds talk in philosophy, but rather the idea that children or adults are supposed to imagine an alternative scenario by making minimal alterations to the scenario that had been presented earlier and identify the counterfactuals that would apply. The reason for this is that, in case talk of possible worlds is brought about, that would score a too easy victory for someone claiming that counterfactual approaches to causation have nothing to do with common sense reasoning with causes.

reason in accordance with Woodward's counterfactual concept of causation. It could be argued, however, that interventionist counterfactuals should be associated with the adult concept of causation.

With respect to the relation between the interventionist concept of causation and cognitive development data, one issue to look into here is whether this move can be made through the use of Woodward's three stages of causal understanding and the experimental evidence from the Bonawitz et al. (2010). As pointed out earlier, the context in which this distinction works involves children's active use of interventions. It could be argued that if children are at the egocentric or agent-centered stage, their understanding of causal relations may not be as objective as to require the understanding of counterfactuals. However, Woodward's interpretation of the data from Bonawitz et al. attributes the fully causal point of view to preschoolers, while Rafetseder et al. (2010) show that preschoolers too have trouble in reasoning counterfactually. Thus, from a psychological point of view the fully causal stage of causal understanding seems to emerge at an earlier time than counterfactual reasoning. Furthermore, it is questionable that the fully causal viewpoint necessitates counterfactuals. Bringing about an event after seeing the event brought about in an agent independent manner may require some kind of mental simulation, but it need not be necessarily tied to counterfactual reasoning. It might as well be the case that the fully causal viewpoint, while accommodating instances of causation that do not involve agents, may be modeled on a concept of causation as agency, in line with the principle of analogical reasoning by Menzies and Price. While one may stand by Woodward's criticism concerning the objectivity of causal relations, the agency concept of causation might work just fine in developmental context. In this particular context, its advantage is that it does not require an understanding of counterfactuals for making agent-independent causal claims. Thus, the evidence against counterfactual reasoning in young children might be taken to support a less objective, albeit agency-based concept of causation.

Bringing into discussion the second constraint that I identified in the previous chapter, the problem for an account such as the one by Woodward is that if causal reasoning is developmentally prior to counterfactual reasoning, the commitment to an objective concept of causation is undermined. However, instances of successful causal reasoning may be explained through a concept of intervention where counterfactuals do not play an essential role, such as the one by Menzies and Price. Counterfactuals along with Woodward's full-fledged concept of causation as intervention may come into the picture later on, in adult causal reasoning. Although made in a different context, I believe that part of the Gijbers and de Bruin argument for the agency concept of causation as a precursor for the interventionist concept of causation[65] supports the idea that the agency concept of causation is developmentally prior. In order to support this claim, the authors also make use of empirical data. While the overall point made by Gijbers and de Bruin is meant to defend the interventionist account against the genesis problem, I believe that one of its implications is that if one is to look for a concept of causation that comes in earlier in development, then the agency concept may be more adequate than the interventionist one.

With this section, I will end the discussion of causation as manipulability in developmental context. One main conclusion to draw is that, while there is evidence of causal reasoning, and particularly intervention-based causal reasoning in young children, it is difficult to accommodate this with Woodward's concept of intervention: for one, children cannot seem to make use of intervention for inferring causal structures in a three variable system; on the other hand, there is evidence against attributing young children the capacity of using counterfactual reasoning, and thus, even if intervention may have a role in young

[65] As Gijsbers & de Bruin point out, the continuity between the two senses of causation is both methodological and conceptual. (p. 1783)

children's causal reasoning, it might not be Woodward's version of interventionism specifically. Another conclusion to draw is that the complexity of causal inference and the methods used for inferring causal claims can be accounted for in a satisfactory manner by adding a temporal component to the manipulationist framework. Another question that deserves further investigation is whether a concept of causation as agency along the lines of the Menzies-Price theory might be less subject to the second constraint mentioned above and, thus, be a more adequate concept of causation for the experimental framework.

The next question to address is whether Woodward's theory provides an adequate account of the adult concept of causation. In the next section I will explain how the non-backtracking interpretation of counterfactuals inherent to Woodward's view might pose problems to everyday uses of counterfactuals in causal contexts. Furthermore, in line with the functional account of causation, I will show that this problem is also present in the context of scientific explanation.

4.4 Interventionist counterfactuals and backtracking

I will now look into the issue of backtracking counterfactuals from the perspective of causation as a functional concept. I will discuss two examples where backtracking counterfactuals can yield into correct judgments regarding the causal structure. The first domain will be everyday causal reasoning, where I will rely on empirical studies into counterfactual questions in causal systems. The second domain concerns causal explanation and counterfactuals in the social sciences, where backtracking counterfactuals have been shown to help in constructing historically plausible causal scenarios. Although none of these cases explicitly relate to Woodward's approach, I will argue that the critiques that can be raised against a Lewis-style analysis of counterfactuals could just as well apply to Woodward's account. The main claim that I wish to argue for in the present section is that, as specified previously, Woodward’s account is too restrictive with respect to cases of backtracking.

4.4.1 Interventionist counterfactuals and backtracking in psychological context

The problem I will address in this section concerns how the relation between causal concepts, counterfactuals and intervention arises in everyday causal claims. I will start by presenting two Bayes networks approaches to counterfactuals and showing that they are supported by empirical research into people’s answers to counterfactual questions in causal contexts. Starting from this data, I will argue that if one assumes that counterfactuals or causation are to be understood through a formal model such as the ones described in this section, or the ones described in chapter 1, there are two stances that one could take towards the truth value of counterfactuals. Opting for one of these stances could result in different conclusions about the graph structure, and, more important for the point that I will try to make, lead to different verdicts concerning which counterfactuals come out true. Finally, I will argue that these results show that Woodward’s version of interventionism about causation only tells one side of the story in relation to causal reasoning and counterfactuals.

4.4.1.1 The pruning and minimal-network models

I will proceed by explaining the formal models that constitute the background for both the empirical studies, as well as some approaches to causal inference. One thing to note here is that both of these approaches seek to define counterfactuals by means of systems incorporating causal relations. While I will not take a stance concerning the issue whether counterfactuals should be understood in terms of causal structures or the other way around, I will be interested in the common points between approaches making use of such models in order to explain causation, counterfactuals, or both.

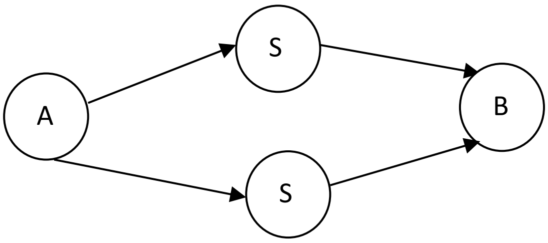
Using the terminology of Rips’s (2009) investigation, two approaches to counterfactuals can be distinguished:

a) The pruning theory, supported by Pearl (2000). According to this theory, the counterfactual state of a variable is evaluated through cutting the connection between the variable in question and its cause(s). It is also very close to the formal basis of Woodward's concept of intervention[66].

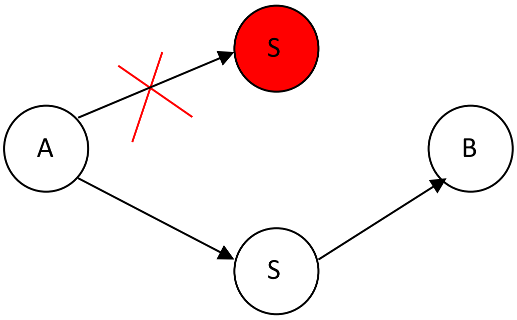
b) The minimal-network theory, supported by Hiddleston (2005). By contrast to the pruning theory, the idea here is to minimize the changes made to the structure of the system. Thus, assuming a counterfactual state (a variable taking a value different from its actual one) implies that the value of its direct cause has been changed, rather than that the causal connection between the respective variable and its direct cause has been severed.

Without going into much detail about these two approaches, I will illustrate the way in which they work with counterfactuals. In order to show how the two models yield into different results, I will choose an overdetermination example, analogous to the one by Gerstenberg et al. 2013. Let us suppose that a commander orders two demolition squads (S1 and S2) to bomb a strategic target. The quantity of explosives used by each squad is enough to cause the building to blow up, but because the mission is considered decisive both squads will deploy the explosives. The commander gives the order, both squads deploy the explosives and the building is destroyed. A counterfactual question here would concern whether the explosion would have taken place if S1 hadn't deployed the explosives. Let us assume the causal system looks like this (A is the commander's order, S1 and S2 are the two squads successfully deploying the explosives and B is the explosion taking place):

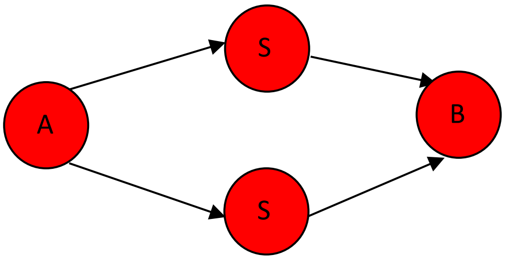
[66] Although both intervention and counterfactuals play an important role in Woodward's theory, as pointed out by Rips in a footnote, it is not clear which explains which, and how circularity can be avoided. I will say more about this in subsection 4.4.1.3. Another interesting thing to note is that, as it will become clear in the next few paragraphs, the pruning theory yields results similar to Lewis's treatment of counterfactuals in causal contexts. It should be emphasized that despite the similar results in the truth value of counterfactuals, the project runs in the opposite direction, namely, explaining counterfactuals through causal systems. A similar treatment of counterfactuals is endorsed by Frisch as well, as pointed out in chapter 2.



According to pruning theory, in case S1 does not occur, that happens by virtue of the connection between A and S1 being severed. Because severing the connection between A and S1 means that A still occurs and causes S2, B ends up occurring.



According to the minimal-network theory, if S1 does not occur, one has to look for its cause somewhere within the system. In the given case, the only way of rendering S1's value to 0 without affecting the graph structure is through A. Since A does not occur, S2 does not occur either, and consequently B does not occur.



Obviously, these are different approaches to counterfactuals and in the current example the consequences for the inferred graph structure are decisive. In what follows, I will mostly be interested in their treatment of backtracking counterfactuals. The pruning theory rules out backtracking counterfactuals. As shown earlier, assuming that S1 does not occur does not imply that A should not occur. In the case of the minimal-network theory, however, some backtracking counterfactuals come out true, that is, given the current graph, seeing that S1 does not occur implies that A does not occur. The backtracking counterfactual 'If S1 had not occurred, A would not have occurred' is true because the only way of making S1 not occur while making minimal changes to the system is through A.

I will now look into an empirical study on how people answer counterfactual questions in causal context and how their answers relate to these two models.

4.4.1.2 Empirical studies into counterfactuals

In a study by Gerstenberg, Bechlivanidis and Lagnado (2013) subjects were asked counterfactual questions with respect to a system with a structure similar to the one described above. It is important to note that, rather than having a real world example, the variables were presented as components of a mechanism that are necessary to activate other components (e.g. A has to function to activate S1 and S2, either S1 or S2 need to function to activate B).[67] The questions are manipulated such that some subjects are asked first about the cause of the counterfactual state (e.g. what would A's value be if S1's value were 0?) and others are asked first about the effect of the counterfactual state (e.g. what would B's value be if S1's value were 0?). This is to see whether the order in which the questions are asked determines results in line with one or the other of the two formal accounts presented above.

The results from previous studies are mixed. For instance, in a study by Lagnado and Sloman (2005), the answers correspond to the predictions of the pruning theory. In a study by Rips and Edwards (2013), however, the results are more in line with the minimal-network theory. As the authors point out, the order in which people answer questions (as in, questions

[67] Gerstenberg et al. use a different notation for the variables. For simplicity's sake, I will use the notation I used for the examples in the previous section.

about the cause of the counterfactual state first, and then questions about its effect, or the other way around) is a crucial difference between the two studies.

Gerstenberg et al. did two experiments. The former replicates the findings of Rips and Edwards, where people are presented with a mechanism where a variable does not work. They were supposed to choose the values of the other variables in the mechanism and also indicate the order in which they ascribed values to the variables. The majority of participants answered in line with the minimal-network theory, looking at the variables on the left first.

In the second experiment, different groups of participants were presented with different causal structures and asked counterfactual questions. The structures were disjunctive (either S1 or S2 is necessary to activate B), or conjunctive (both S1 and S2 are necessary to activate B). The assumptions about the state of functioning of the system also differed. In one case it is assumed that all components were operating and participants are required to say what would happen if S1 were not operating. In the other case, it was assumed that none of the components were operating and participants were required to say what would happen if S1 were operating. In all these cases, the order in which participants were asked the question differed. Assuming that S1 does/does not operate, different groups of participants were asked about the other components either in the order A-S2-B or B-S2-A.

The results showed that the answers differ when the order of the components differs. Thus, in the A-S2-B condition, out of the total of 160, 36 participants answered in line with the minimal network theory, and 42 in line with pruning theory. In the B-S2-A condition only 15 participants answered in line with minimal network theory and 68 consistent with pruning theory. Another interesting result, not predicted by either of the two theories, was that participants were less certain of A’s value than of B’s value. The authors’ explanation is that people may process counterfactual questions locally, rather than in relation to the whole system. Upon deciding on the value of B or S2, they may be confronted with an inconsistency when deciding A’s value and have to choose between A working and S1 malfunctioning, or A not working and S2 working spontaneously.

The authors conclude that subjects are more in line with the minimal-network theory when asked about causes, and more in line with pruning theory when asked about effects. They argue that the overall pattern shows that people process counterfactuals in a local manner, rather than taking the whole system into consideration.

The authors also point out that a direction in which further research could be done would concern asking counterfactual questions in connection with less abstract systems. I believe that this is the right way to go if one is looking into causal structures and counterfactuals from the perspective of everyday reasoning. On the other hand, I will shortly argue that once real world examples are used, it becomes even harder to distinguish which one of the possible formal approaches is at work. I will now go on to investigate how these findings could be relevant to an interventionist take on causation and come back to this last point afterwards.

4.4.1.3 Bringing interventionism into the picture

Where does intervention come into place given the previous discussion? According to the formal models presented above, counterfactuals are understood through a causal system where one variable is assumed to gain a value different from its current one (i.e. if all components in the system are supposed to operate, that variable does not operate). The pruning theory has a very specific account of how a variable’s value is set in a causal system: an intervention through a variable outside of the system that cuts the connection between the variable intervened upon and its parents. Although the minimal-network theory does not explicitly discuss intervention, it can be inferred that intervening to change a variable while minimally altering the structure of the graph amounts to changing it through its cause.

Woodward's (2003) theory shares the formal structure presented by the pruning account. However, as Rips (2010) points out in a footnote, it is difficult to make sense of counterfactuals through the Bayes nets apparatus from the perspective of Woodward's view because Woodward defines interventions by means of counterfactuals. As shown earlier, counterfactuals are needed in order to account for an interventionist understanding of unmanipulable causes. Furthermore, Woodward's definition of intervention employs causal terms. We are thus presented with the use of cause to define intervention, and the use of intervention to define causation and counterfactuals. As specified in chapter 1, Woodward acknowledges the circularity, but denies it being vicious.

Although the ties between Woodward's idea of manipulability and Pearl's account of causal inference are obvious, it should be noted that these problems do not arise in the case of Pearl's account. By taking causation as a primitive, Pearl can go on and use causal relations to analyze interventions and build a model of counterfactuals from the same framework. Given this picture, one could go on and say that Pearl's endeavor is the opposite of what we see in Woodward. One can construct a formal model of how interventions and counterfactuals work by taking causation as a primitive, but can one also use the former two in order to explain causation?

The issue seems to be even more complex given that Woodward does not seem to opt for an understanding of counterfactuals independent from his interventionist claims. As previously pointed out, the formal framework of Woodward's account seems to be similar to Pearl's pruning account. Since this approach results in similar truth values for counterfactuals as Lewis's approach, I believe it to be a further ground for the similarity between Woodward's and Lewis's theories. The fact that Lewis' embarks on a different project is part of the divergent aims as far as metaphysics is concerned. The reason why I take Pearl's approach to counterfactuals, rather than Lewis's to be closer to Woodward's account is that

Woodward rejects Lewis's similarity criteria and does not seem to endorse the talk about possible worlds.

Since I will not be focusing on solving this conceptual tangle here (if it can be solved in any way, that is), I am only going to point to a potential way of making sense of Woodward's version of interventionism and its relation to the previously discussed empirical studies. Thus, leaving the metaphysical concerns aside, and subscribing to the functional view discussed in chapter 3, one could consider a cluster of concepts containing intervention, counterfactuals and causation without seeking to define them in terms of concepts outside of this cluster. The cluster is at use when people make causal inferences. One can understand causation in terms of intervention and counterfactuals just as well as one can understand counterfactuals in terms of interventions on causal variables in a system. In conformity with the presented formal approaches, causal understanding takes place within a directed system of variables with given connections and probability relations. This does not shed much light on the deeper philosophical issues, but it may point to a way of making sense of the relation between causal inference and counterfactuals in everyday reasoning.

The aspect that I am going to concentrate on, as far as Woodward's account is concerned, is that backtracking counterfactuals need to come out false. Because counterfactuals are analyzed in terms of interventions that keep the causes of the variable intervened upon fixed, a counterfactual state is assumed to have occurred through a separate intervention, rather than through changing the values of other variables in the system.[68] Thus, on Woodward's account only forward looking counterfactuals can be true. Woodward also relies on empirical data on causal reasoning pointing out to such results: 'these results seem inconsistent with claims (e.g., Bennett, 1984) in the philosophical literature that people either

[68] I emphasize that I am talking about changes in a system, and not about the idea that effects can be changed through their causes, but not the other way around. As stated earlier, I do not believe that on Woodward's account this idea can be used to explain why causal relations are asymmetric, since he is using the concept of causation in defining intervention.

do not distinguish at all between backtracking and nonbacktracking counterfactuals or do not preferentially employ the latter in contexts involving causal reasoning.' (Woodward 2007: 29) However, as the studies of Rips and Edwards, and Gerstenberg et el. show, people sometimes do use backtracking counterfactuals in causal contexts. The case presented in the Gerstenberg study shows that, if one is given a specific value of a variable and asked about its cause, one may look for its immediate cause within the system and ascribe it a corresponding value, rather than opting for cutting the arrow between the given variable and its parent. Based on this backward-looking view, the other variables in the graph are assigned distinct values. Note that in an overdetermiantion case as the one presented above, opting for a forward or backward-looking view yields into completely different views on the values that the variables in the graph end up taking. Thus, to come back to Woodward's claim, according to the studies discussed above, people prefer to use nonbacktracking counterfactuals in causal contexts where they are asked about the effect of a counterfactual state, but not where they are asked about its cause.

Once this broader point of view is taken into account, one can contend that causes can be inferred from known effects, and effects can be inferred from known causes. Both kinds of information are available within a given graph structure. Suppose that one is presented with a device with four components where component 1 activates both components 2 and 3, while either of the components 2 or 3 is needed to activate component 4. Given that no information outside the structure of the system is given, if one only knows that 3 is working, one can accept that the following counterfactual is true: 'If component 3 had not worked component 4 would still have worked.' However, someone knowing that component 3 works can also endorse the following counterfactual: 'If component 3 had not worked, component 1 would not have worked either'. For an approach such as the one by Woodward, the problem is that through leaving the metaphysical issues aside (namely, the fact that causation must be an asymmetric relation, and so must be counterfactuals if one is opting for a counterfactual analysis), counterfactuals can work both ways. Opting for one view or the other leads to different value distributions in the graph, and it all seems to depend on which variables one considers first.

I believe that this data brings about an important point concerning the investigation of causation and counterfactuals through formal models. Given that the model contains enough information to allow for different conclusions, people choose to answer counterfactual questions in accord with the elements on which they focus their perspective: causes or effects. Given the value of a variable in a system, one could take:

a) A backward-looking stance, focusing on what determined a particular variable's value.
b) A forward-looking stance, focusing on what would be the changes in the effect variable given a certain value of its parent variable.

This shows that Woodward's concept of intervention tells only the forward-looking part of the story. As long as the discussion takes place in the context of causal reasoning and inference, the backward-looking stance is not incompatible with an interventionist view on causation. Unlike in Woodward's version of interventionism, changing a variable's value can be done through changing its cause. I also emphasize that this concept of intervention is compatible with the way one may think of intervening on an everyday basis. Suppose for instance that someone is experiencing a headache while also knowing that most of her headaches are caused by too much blood flow. If she knows that caffeine reduces blood flow, she can intervene on her headache by drinking a cup of coffee. I believe this a good example of intervention where one can change a variable through intervening on its direct cause. It involves detailed knowledge of the workings of the system, but so does Woodward's concept of intervention.

The problem I wish to point out here is that Woodward's (or any other account committed solely to the forward-looking view) lacks the generality that is required to account for the way in which people reason with causation and counterfactuals. In the light of the issues discussed in the previous chapters, it is perfectly understandable why Woodward would endorse such a view: his theory has aims beyond causal inference. In that case, however, he would also have to commit to a metaphysical claim holding that only forward-looking counterfactuals would constitute a suitable analysis of what causation is. This would mean that on a pragmatic level both views are acceptable, while only the forward-looking view explains causation on a fundamental, metaphysical, level. Note that this solution would bring along problems of its own, such as explaining why it is necessary to have two distinct ways of understanding counterfactuals at the metaphysical and epistemic level. While I will not pursue this issue further, I would like to point out that it is another case that illustrates the tension within Woodward's account: constraints that should apply to a metaphysical account are exercised upon causal inference.

Coming back to the experimental data now, there may be a worry about how much support the backward-looking stance has in the Gerstenberg et al. study. After all, even in the A-S1-B condition, there are more answers in line with the pruning account (42 versus 36 out of 160). It should be noted, though, that even if there are more answers in line with the pruning account on this condition, there is a significant number of answers that involve backtracking. Accepting a backward-looking stance as a valid take on counterfactuals and causal reasoning can explain why a significant number of participants answered in line with the minimal-network theory. Another thing to note is that taking the answers that pertain to a backward-looking stance and those that pertain to a forward-looking stance together yield into a bit less than 50% of the answers. That means that slightly more than half of the participants had answers that were not predicted by either of the two formal accounts. An interesting question here would be whether the participants were simply confused about what to answer, or whether there could be a different way in which they make sense of how counterfactuals work in causal systems.

The second worry concerns the abstract nature of the task. And the issue here would be what to look for when considering more common uses of counterfactuals in causal reasoning tasks. From the point of view of a study that investigates the predictions of the two formal accounts, the fact that the participants were presented with variables as parts of a mechanism may count as an advantage. There is no ambiguity concerning what variables to include in the system or how many values they can take. In a more complex scenario it is much harder to single out all the relevant variables, the values they could take and the causal connections.[69]

4.4.2 Backtracking and counterfactuals in the social sciences

In this section I will go through some uses of backtracking counterfactuals in the context of causation, causal inference and explanation in the social sciences. I will rely on Reiss's (2009, 2012b) work, mainly on counterfactuals and explanation in history.

Reiss (2012b) tackles the issue of counterfactuals in the social sciences first by pointing out their connection to causation and causal inference and by presenting Lewis's semantics. As specified previously, Lewis's semantics of counterfactuals yields into similar results with respect to their truth values as the above-mentioned formal account by Pearl. Further, Reiss discusses the desiderata of social scientists with respect to counterfactuals. Two of them will be relevant for the problem that I am discussing here.

The first is cotenability, that is, 'whatever else we assume in order to make the counterfactual true should not be undermined by the counterfactual antecedent' (Reiss 2012b:

[69] See also Hall's (2007) discussion of the problem of variable choice.

161) He makes use of a similar example as Lewis (1973), taking the following sentences: A – Jim asks Jack for a favor; B – Jim and Jack quarreled yesterday; C – Jack does not grant Jim the favor. However, we know that Jim is very prideful and he would never ask Jack for a favor after a quarrel, so if Jim asked Jack for a favor, Jack would grant it. According to Lewis's resolution, the counterfactual 'If Jim had asked Jack for a favor, Jack would not have granted it' is true since according to the Lewisian semantics the backtracking counterfactual 'If Jim had asked for a favor today, he would have not quarreled with Jack yesterday' comes out as false.

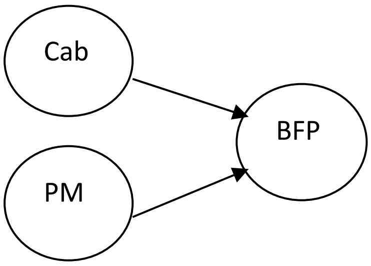
Taking a look at the goals of social scientists, Reiss points out that they 'aim to keep as much as possible about historical actors' situations and dispositions intact.' and that 'in order to achieve cotenability (…) counterfactuals will sometimes have to backtrack' (Reiss 2012b: 162). Whether backtracking counterfactuals come true or false depends on the kinds of events involved and on the strength of the evidence. For instance, if the quarrel was accidental, the counterfactual 'If Jim asked Jack for a favor, Jack would grant it' could come out true. The backtracking counterfactual 'If Jim had asked for a favor today, he would have not quarreled with Jack yesterday' would also come out true. However, if the quarrel is an important event, the result of a tense friendship between Jim and Jack, the counterfactual 'If Jim had asked Jack for a favor, Jack would not have granted it' would come out as true.

The second desideratum relevant to my investigation is historical consistency. Reiss uses an example from the historian Yuen Foong Khong (1996) who asks whether World War 2 could have been avoided if the UK foreign policy had been more confrontational. As Reiss points out, 'a Lewis counterfactual would make the antecedent true by miracle: a surgical intervention that changes nothing but the UK foreign policy'[70] (Reiss 2012b: 163). The problem with this, Reiss notes, is that it would violate what we know about the UK

[70] Note how this description could just as well apply to Woodward's concept of intervention!

government back then. We know that Neville Chamberlain was prime minister at the time and that his policies were meant to avoid war at all costs. A more consistent historical scenario would involve backtracking: if the UK foreign policy had been more confrontational, someone else would have had to be prime minister. Thus, one could build a scenario in which the UK, lead by Churchill rather than Chamberlain, would have lead a more aggressive foreign policy and avoided World War 2.

The way in which Reiss's approach to counterfactuals in the social sciences connects with the previous discussion is through his use of Hiddleston's causal model for counterfactuals. The cases presented earlier could be interpreted through the framework proposed by Hiddleston. Let us take into consideration a graph containing the British prime minister, the members of the cabinet and the UK foreign policy. If a disjunctive structure is assumed, in order to set the policy to confrontational from pacifist, either the prime minister has to be set from Chamberlain to Churchill, or the members of the cabinet from pacifist to confrontational. This sort of intervention is covered by the Hiddleston account. By contrast, a Woodward-style intervention would involve leaving the prime minister and cabinet members variables intact and alter the UK foreign policy separately.

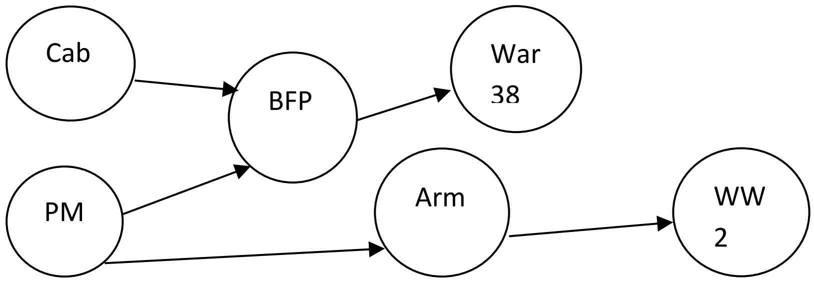


While Reiss discusses several problems[71] for such an approach to counterfactuals, I will only discuss the issue of backtracking. On Lewis's approach, and on approaches such as the one by Pearl, or Woodward, that rely on the same idea that backtracking counterfactuals

[71] I will not discuss circularity since Reiss's talks about it in the context of causal inference. As I mentioned earlier, I do not think that using causal relations to understand counterfactuals or going the other way is problematic when inferring causal relations.

must turn out false, the problem with backtracking is that it could lead to wrong claims with respect to causal structure. For instance, considering that had the symptoms not been present, a certain disease would not have been present could lead one to think that the symptoms cause the disease. Reiss points out that in cases of backtracking, counterfactuals are no longer reliable indicators of the causal order. In these cases an intervention could tell which causes which: while treating the disease makes the symptoms go away, treating the symptoms does not make the disease go away.

In the previously discussed historical examples, however, intervention is impossible. While using a Lewisian semantics would lead to historically inconsistent causal judgments, backtracking could also lead to counterexamples. I will make use of the previous historical example and a simplified version of the graph discussed by Reiss.



The graph contains an additional plausible causal path, where Chamberlain not being prime minister causes earlier rearming for Britain which in turn causes World War 2. In this structure one can take World War 2 to counterfactually depend on British foreign policy, even though in this case it is not BFP that affects its likelihood. As Reiss points out, there is a dilemma here: either accept Lewis-style counterfactuals and end up with a historically implausible scenario, or use backtracking and end up with wrong causal claims.

The solution proposed by Reiss involves taking causal background knowledge into account. He goes on to claim that 'Lewis is mistaken to call the nonbacktracking resolution of vagueness ordinary or standard; it is just one resolution among others. In fact, there are good reasons to believe that ordinary language counterfactuals standardly backtrack' (Reiss 2012b: 174). The everyday cases of backtracking involve evidential reasoning[72]: while the symptoms do not cause the disease, they constitute good evidence for the presence of the disease.

Reiss further proposes to amend the minimal network account through specifying that the minimal model 'may contain only variables that are connected to the putative effect variable, if at all, only through directed paths that include the putative cause variable' (Reiss 2012b: 174-175) Thus, on a path that does not contain the putative cause variable, the backtracking counterfactual is not acceptable. Coming back to the previous example, upon investigating whether BFP could have prevented WW2, one needs to rule out the causal path PM -> Arm -> WW2, because it does not contain the putative cause (BFP).

4.4.3 Backtracking, intervention, and temporal direction

Having discussed two areas where backtracking counterfactuals can be of use, the conclusion to draw from a functional perspective is that Woodward's approach to counterfactuals inherent to his theory of causation may be too restrictive. Nevetheless, a defeasible claim to make would be that Woodward's concept of intervention can account for a part of the uses of causal claims and counterfactuals. However if this is to be taken as the standard view on which counterfactuals come to be true, there may still be a question about why models that accept backtracking are successful.

A question to ask is whether there can be a distinction between two understandings of causal models of counterfactuals, analogous to the distinction between two concepts of causation. These two ways of understanding counterfactuals through causal models could be taken to be useful in different contexts. Since my interest here is in causation and some of its

[72] Reiss discusses some examples; I will not focus on them right now, since I believe that the discussion in the previous sections of this chapter made it quite clear that ordinary causal judgments sometimes involve backtracking.

features at use in functional perspective rather than causal models of counterfactuals, I will not explore this issue here. One point that I wish to make, however, is that there is a way in which the pruning and minimal-network models seem to be irreconcilable: the possibility of having true backtracking counterfactuals. From this perspective, they cannot be seen as the distinction between causation as production and causation as difference-making. Whereas Hall (2007) sees the two concepts as complimentary, this does not hold about the two causal models which have conflicting assumptions about the truth values of counterfactuals. Thus, as far as causation is concerned, if counterfactuals are understood through the pruning model (as is the case with Woodward's account) the usefulness of backtracking counterfactuals in functional contexts is not only left unexplained, but may also be counted as a counterexaple to one of the main tenets of such theories.

Another question that I would like to answer here concerns the motivation of ruling out bactracking counterfactuals. One thing to note is that a model such as the ones at use in Pearl's or Woodward's approaches to causation seems to require a non-backtracking interpretation of counterfactuals. Nevertheless, a functional account of causation and one based on manipulability particularly need not be tied to a single model, as suggested above, there are possible interventions on a minimal-network system as well. Furthermore, some of the evidence concerning everyday counterfactual reasoning my suggest that, at least in these contexts, people might not be working with systems of variables at all, but rather looking for causes or effects of certain variables.

In line with my earlier claims, I believe that opting for an exclusively non-backtracking interpretation of counterfactuals can also be taken as an expression of the tension between metaphysical and functional considerations on causation in Woodward's account. While, as discussed in chapter 2, there may be other ways in which Woodward's version of interventionism can explain the asymmetry of causation, the non-backtracking interpretation of counterfactuals is one of them. In relation to my claims from chapter 2, I would like to point out that if the direction of causation is given by the direction of time then no commitment needs to be made concerning the falsity of backtracking counterfactuals. If temporal direction is taken to be among the features of causal relations that ground manipulability, or if causal understanding is taken to be more fundamentally related to temporal order, then there need not be any special interpretation of counterfactuals that should match the causal asymmetry. One can make sense of causal notions through counterfactual models, and if the causal notions are already taken to be asymmetric, then one need not limit the way in which counterfactuals could be understood. Of course, such a solution would also work if causal relations are taken to be fundamentally asymmetric, or if the causal asymmety is accounted for in some other way, as long as the solution is does not involve some features of a particular causal model.

To conclude this chapter, I have been illustrating how Woodward's version of manipulability has some shortcomings in functional contexts. I have mostly looked at data from psychology where both the complexity of Woodward's approach to intervention as well as the different kinds of evidence at use in causal inference prove that the interventionist concept of causation might be, at best, accounting for a part of the uses of causal concepts in causal learning and causal reasonig. Furthermore, in some cases, the uses of interventions may not be in line with Woodward's specific definition and other concepts, such as that of agency could be applied. Finally, I have shown that both in an empirical as well as a philosophy of science context, the standard reading of counterfactuals need not completely rule out backtracking. This also shows the limitations of Woodward's account. Finally, I have pointed out that in certain cases the project of connecting manipulability to temporal order might be of help even in a functional project. For one, it would fit in with the developmental data on the prevalence of the use of temporal cues in inferring causal structures. Secondly, because it would take care of the causal asymmetry problem, it would not be necessarily tied to a non-backtracking interpretation of counterfactuals. To connect the discussion in this chapter with the previous remarks on the interactions between a metaphysical project and a functional project and to the asymmetry issue, it is worth pointing out that solving various issues faced by the interventionist concept of causation can be done in two ways. This amounts to the distinction between the objective and the subjective view discussed in chapter 2. The subjective package would contain the claim that the agent's perspective determines the arrow of time as well as the arrow of causation and also the consideration that the interventionist concept of causation is developed from an agency one. The objective package would connect manipulability to temporal direction in order to explain the direction of causation and also admit of the connection between causation and temporal sequence as being developmentally prior to the one between causation and intervention. Someone sticking with a realist project would find the objective view more appealing.

Conclusions

I will end by summarizing my approach to the several problems discussed in this dissertation. The first problem concerns manipulability as a metaphysical approach to causation. As I pointed out in chapter 1, in the relevant literature (namely Woodward and Menzies and Price) metaphysical and conceptual claims are sometimes mixed. I have argued that there is good reason to look into manipulability as a metaphysical approach to causation. One of the points I made is that if the usefulness of manipulability-based causal claims is explained by people successfully tracing ‘worldly features’ of causal relations, then an inquiry into these features is necessary. While I do not believe this to be a shortcoming of Woodward’s theory as long as it is read as a conceptual account of causation, I believe that it is an issue worth investigating. Particualrly, I have been looking at some of the issues that a manipulationist theory committed to causal realism would need to face. I have, thus, started by looking into reasons why one would choose causal realism over projectivism from a manipulationist perspective. I have argued that in the versions of Woodward and Menzies and Price, both accounts can reach an acceptable degree of objectivity, namely in the sense that the truth value of causal claims need not vary with different agents. For this particular aim, Woodward’s version of manipulability is not necessarily preferrable. Particularly, I have pointed to a problem that might raise questions over the fit between Woodward’s account and causal realism: the dependence of causal concepts on systems of variables, the choice of which is ultimately decided by the agent. However, I have argued that there is one reason why one may prefer realism over projectivism. This goes back to the argument in the fashion of the ‘No Miracle’ argument: explaining why the success of manipulability-based causal claims may come down to pointing to a correspondence between the features of causal relations in the world and the features that people exploit for manipulation and control purposes.

From the same metaphysical perspective I have been looking at an issue that objectivist approaches to causation as manipulability are confronted with: explaining the asymmetry of causation. In this respect, I have pointed to an argument by Mackie against explaining the asymmetry feature through manipulability. The problem is that manipulating the putative cause variable makes use of a different causal relation, which in turn is supposed to be asymmetric. While from a projectivist perspective this problem can be dealt with if the perspective of the agent is taken to be prior to causal or temporal reasoning, the causal realist would have to have an account of what makes causal relations asymmetric. In investigating possibilities of answering this problem from the perspective of Woodward's approach, I have argued why reliance on counterfactuals is unsatisfactory and explained why taking causal relations and their asymmetric feature as primitive would not answer the worry of the philosopher interested in metaphyics. I have investigated some ways in which the causal asymmetry can be accounted for and pointed to some of their weaknesses. I have presented a potential solution: linking manipulability to temporal direction. From a metaphysical point of view, my solution involves breaking the higher-level manipulability relation into the features that ground it. This is coherent with the way in which Woodward argues for his version of manipulability being tied to realism. My claim is that temporal direction is to be counted among the worldly features that ground manipulability. This way, one can make sense of the higher-level causal claims based on manipulability and the way in which they make use of an asymmetric relation rather than be taken as grounds for the causal asymmetry. I have proceeded to answer typical objections against equating the causal and temporal asymmetries. Notably, I have argued against backwards causation and I contended that simultaneous causation is possible on higher-level instances, but temporal order comes into place when the causal relation is broken down into more fundamental entities. Finally, I hold, as a weaker claim, that causal asymmetries and temporal asymmetries can be connected at least at the level of causal understanding. This claim has been useful for the second part of my thesis, where I have investigated the manipulability theory of causation as a functional project.

There are a couple of ways in which my former, metaphysical investigation connects to the latter, functional one. One of them involves precisely the argument enounced above about the connection between a philosophical theory about causation and its uses in psychological or more general scientific contexts. Another one concerns the asymmetry problem discussed in chapter 2: in psychological context there is evidence for an asymmetry between causal and diagnostic reasoning. Even though one might not care about metaphysical foundations, from a functional perspective it may be helpful to explain this asymmetry. My claim is that temporal judgments are closely connected to causal judgments, being the source of the asymmetry in causal reasoning. Experimental evidence shows that interventions work better when they are accompanied by temporal cues, yielding into successful causal inference. Finally, I have argued for a perspective where some of the functional and metaphysical types of claims are intertwined and influence each other. I have thus proceeded to single out some features of Woodward's version of manipulability that may be explained through tacit metaphysical assumptions and that result in limiting the applicability of manipulability to experimental and philosophy of science contexts.

The first constraint concerns the definition of causation as manipulability and the more or less 'verificationist' (Reiss 2008) stance on causal claims and evidence. I have illustrated how the geometrical-mechanical account of causation does a better job in accommodating temporal evidence (as shown in Waskan 2012) and also that there are problems with using interventions in causal inference (Frosch et al. 2012). I have further argued that connecting manipulability to temporal direction (as in Lagnado and Sloman, 'timely intervention') may increase the degree of generality of the model and the range of evidence used. The second constraint concerns the use of counterfactuals to spell out the fully agent-independent concept of causation. Once again, I argued that this is problematic in developmental context where there is evidence that children employ conditional rather than counterfactual reasoning (as shown by Rafetseder et al.). The third constraint concerns the exclusion of true backtracking counterfactuals. Based on evidence from everyday reasoning (Gerstenberg et al.) and a philosophy of social sciences example (Reiss 2012), I have pointed out that the understanding of counterfactuals inherent to Woodward's account is at best limited. I have further argued that accepting different causal models that permit some backtracking counterfactuals would not be problematic if the asymmetry of causation is understood through temporal direction (or in a different way than an exclusively non-backtracking version of counterfactual dependence).

There are some more general remarks about these issues that could be made at this point. One is that the possibility of supporting a manipulability approach to causation along with causal realism needs more investigation and that this investigation is to be made mostly on metaphysical grounds. I have pointed out that an objectivist view on manipulability and causal relations needs to provide a satisfactory account for the source of causal asymmetry. I have also emphasized a possible reason to doubt that Woodward's definition of causal concepts may fit in with the project – the dependence of direct causation and contributing causation on the choice of variable sets. While there are reasons to pursue such project, I believe that the quest for a fully agent-independent concept of causation as manipulability is still unresolved. Another remark I wish to make is that I agree with Woodward's claim that success on causal reasoning tasks is best explained through correspondence with causal relations in the world, and that I believe this to make a strong case for realism. However, unlike Woodward, I take this to be an additional motivation to look into the metaphysical foundations even when pursuing a functional account. Finally, from a functional point of view I believe there is much work to be done concerning the different causal models, their usefulness and the more general, philosophical, perspective that can explain the workings of causal concepts and their connection to manipulability.

Bibliography

Albert, D. (2000) *Time and Chance*, Cambridge Mass: Harvard University Press.

Bayes Nets, URL = <http://www.bayesnets.com/>

Bechlivanidis, C. & Lagnado, D. (2013) 'Does the "why" tell us the "when"?', *Psychological Science* 24 (8): 1563-1572.

Bennett, J. (2003) *A Philosophical Guide to Conditionals*, Oxford University Press.

Black, M. (1956) 'Why Cannot an Effect Precede its Cause', *Analysis*, 16: 49–58.

Bonawitz, E., Ferranti, D., Saxe, R., Gopnik, A., Meltzoff, A., Woodward J., Schulz, L. (2010) 'Just Do It? Investigating the Gap between Prediction and Action in Toddlers' Causal Inferences', *Cognition* 115:104–17.

Cartwright, N. (2002) 'Against Modularity, the Causal Markov Condition, and Any Link Between the Two: Comments on Hausman and Woodward', *British Journal for the Philosophy of Science*, **53:** 411–53.

Cartwright, N. (2004) 'Causation: One word, many things', *Philosophy of Science* 71: 805-819.

Chakravartty, A., 'Scientific Realism', *The Stanford Encyclopedia of Philosophy* (Summer 2013 Edition), Edward N. Zalta (ed.), forthcoming URL = <http://plato.stanford.edu/archives/sum2013/entries/scientific-realism/>.

Downing, P. B. (1959) 'Subjunctive Conditionals, Time Order, and Causation', *Proceedings of the Aristotelian Society* 59(1959): 125-40.

Dummett, M. (1964) 'Bringing About the Past', *Philosophical Review*, 73: 338–59.

Elga, A. (2000) 'Statistical Mechanics and the Asymmetry of Counterfactual Dependence', *Philosophy of Science* 68: 313-24.

Fenker D., Waldmann M.R., Holyoak K.J. (2005) 'Accessing causal relations in semantic memory' Memory and Cognition 33: 1036–46.

Frisch, M. (2014) *Causal Reasoning in Physics*, Cambridge University Press.

Frisch, M. (2013) 'Time and Causation', in H. Dyke, A. Bardon (eds.), *The Blackwell Companion to the Philosophy of Time*.

Frosch, C., McCormack, T., Lagnado, D., Burns, P. (2012) 'Are Causal Structure and Intervention Judgments Inextricably Linked? A Developmental Study', *Cognitive Science* 36: 261-285.

Gerstenberg, T., Bechlivanidis, C. & Lagnado, D. A. (2013) 'Back on track: Backtracking in counterfactual reasoning' in M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Gerstenberg, T., Goodman, N. D., Lagnado, D. A. & Tenenbaum, J. B. (2015) 'From counterfactual simulation to causal judgment', *Proceedings of the 36th Annual Conference of the Cognitive Science Society.*

Gibbard, A. & Harper, W. 1978. 'Counterfactuals and Two Kinds of Expected Utility', in C. Hooker, J. Leach, and E. McLennen (Ed.*), Foundations and Applications of Decision Theory Foundations and Applications of Decision Theory*, D. Reidel, Dordrecht, Holland, pp. 125–162.

Gijsbers, V., De Bruin, L. (2014) 'How agency can solve interventionism's problem of circularity', Synthese 191: 1775-1791.

Glymour, C. (2000) 'Bayes nets as psychological models', in Keil, F. & Wilson, R. (eds), *Explanation and Cognition*, Bradford Books.

Glynn, L. (2012) 'Review of Mumford and Anjum: *Getting Causes from Powers*', Mind 121 (484): 1099-1106.

Gopnik A., Sobel, D., Schulz, L., Glymour, C. (2001) 'Causal learning mechanisms in very young children: Two-, three- and four-year-olds infer causal relations from patterns of variation and covariation', *Developmental Psychology* 37: 620-629.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, D. E., Kushnir, T., & Danks, D. (2004) 'A theory of causal learning in children: Causal maps and Bayes nets' *Psychological Review, 111*, 1-31.

Griffiths, T. L., & Tenenbaum, J. B. (2009) 'Theory-based causal induction', *Psychological Review,* 116, 661-716.

Hagmayer, Sloman, Lagnado & Waldmann (2007) 'Causal learning through intervention', in Gopnik, A., Schulz, L. (eds), *Causal Learning*, Oxford University Press.

Hall, N. (2004) 'Two Concepts of Causation', in Collins et al., *Causation and Counterfactuals*, The MIT Press.

Hall, N. (2007) 'Structural Equations and Causation', Philosophical Studies 132 (1): 109-136.

Harris, P. (2000) *The work of the imagination*, Oxford, England: Blackwell.

Hartmann, S. & Bovens, L. (2003) *Bayesian networks in philosophy.* In: Lowe, B., Malzkorn, W., Räsch, T. (eds.) *Foundations of the formal sciences II: applications of mathematical logic in philosophy and linguistics*, Trends in logic (17), Springer, Dordrecht, 39-46.

Hausman, D. & Woodward, J. (1999) 'Independence, Invariance and the Causal Markov Condition', *British Journal for the Philosophy of Science* 50: 521-583.

Hausman, D. (1998) *Causal Asymmetries*, Cambridge University Press.

Henschen, T. (2015) '*Ceteris Paribus* Conditions and the Interventionist Account of Causality', *Synthese* 1 (2015).

Hiddleston, E. (2005). A Causal Theory of Counterfactuals, Nous 39(4), 632–657.

Hitchcock, C. (1993) 'A Generalized Probabilistic Theory of Causal Relevance', *Synthese* 97 (3): 335-364.

Hitchcock, C. (2011) 'Probabilistic Causation', *The Stanford Encyclopedia of Philosophy (Winter 2011 Edition),* Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/win2011/entries/causation-probabilistic/>.

Hitchcock, C. (2000) 'Review of Hausman, *Causal Asymmetries*', British Journal for the Philosophy of Science 51: 175-179.

Hong L., Chijun Z., Xuemei G., Shan G., Chongde L. (2005) 'The influence of complexity and reasoning direction on children's causal reasoning', *Cognitive Development* 20(1):87–101.

Hume, D. (1975) *Enquiry concerning Human Understanding,* in *Enquiries concerning Human Understanding and concerning the Principles of Morals,* edited by L. A. Selby-Bigge, 3rd edition revised by P. H. Nidditch, Oxford: Clarendon Press, 1975.

Kant, I. (1965) *Critique of Pure Reason,* orig. 1781, trans. N. Kemp Smith. New York: Macmillan Press.

Khong, M.F. (1996) 'Confronting Hitler and its Consequences', in P. Tetlock & A. Belkin (eds) *Counterfactual Thought Experiments in World Politics*: 95-118, Princeton NJ: Princeton University press.

Lagnado D. (2011) 'Causal thinking', in Mckay Illari, P., Russo, F., Williamson, J. (eds) (2011) *Causality in the Sciences* .

Lagnado, D. & Sloman (2002) 'Learning Causal Structure', in W. Gray & C. D. Schunn (Eds.), *Proceedings of the twenty-fourth annual conference of the cognitive science society,* Erlbaum.

Lagnado, D. & Sloman (2004) 'The Advantage of Timely Intervention', *Journal of Experimental Psychology: Learning, Memory, and Cognition* 30 (4): 856-876.

Lewis, D. (1972) 'Psychophysical and Theoretical Identifications', *Australasian Journal of Philosophy* 50: 249–58.

Lewis, D. (1973): 'Causation', *Journal of Philosophy* 70: 556–567.

Lewis, D. (1976) 'The Paradoxes of Time Travel', *American Philosophical Quarterly*, 13: 145–52.

Lewis, D. (1979), ‘Counterfactual Dependence and Time’s Arrow’, *Nous* 15:4 (1979): 455-476.

Lewis, D. (1986) ‘Causal Explanation’, in Lewis*, Philosophical Papers,* vol. II, Oxford University Press: 214-241.

Lewis, D. (2004) ‘Causation as Influence’ in Collins et al., *Causation and Counterfactuals,* The MIT Press: 75-106.

Loewer, B. (2007) ‘Counterfactuals and the Second Law’, in *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Huw Price and Richard Corry (ed.), New York: Oxford University Press, pp. 293-326.

Loewer, B. (2012) ‘Two Accounts of Laws and Time’, *Philosophical Studies* 160: 115-137.

Mackie, J.L. (1965) ‘Causes and Conditions’, *American Philosophical Quarterly* 2 (1965): 245-264.

Marmodoro, A. (2009) ‘Do powers need powers to make them powerful? From pandispositionalism to Aristotle’, *History of Philosophy Quarterly,* Vol. 26, No. 4, pp. 337-352.

Marmodoro, A. (2014) ‘Causation without Glue: Aristotle on Causal Powers’ in C. Natali, C. Viano and M. Zingano (eds.), *Aitia. Les Quatre Causes d'Aristote. Origins et interprétations,* Peeters, Louvain, pp. 221-246.

McCormack, T., Butterfill, S., Hoerl, C., Burns, C. (2009) ‘Cue competition effects and young children’s causal and counterfactual inferences’, *Developmental Psychology* 45, 6: 1563-1575.

McCormack, T., Frosch, K., Patrick, F., Lagnado, D. (2014) ‘Temporal and statistical information in causal structure learning’, *Journal of Experimental Psychology, Learning-Memory and Cognition*.

Mellor, D. H. (1995) *The Facts of Causation*, London: Routledge Press.

Menzies, P. & Price, H. (1993) ‘Causation as a Secondary Quality’, *British Journal for the Philosophy of Science*, **44**, pp. 187–203.

Menzies, P. (1996) ‘Probabilistic Causation and the Pre-emption Problem’, *Mind*, 105: 85–117.

Menzies, P. (2009) ‘Platitudes and Counterexamples,’ in Beebee, H., Hitchcok, C., Menzies, P. (eds), *The Oxford Handbook of Causation*..

Mumford, S., Anjum, R.L. (2011) *Getting Causes from Powers*, Oxford University Press.

Pearl, J. (2000) *Causality*. New York: Cambridge University Press.

Pearl, J. (2001) 'Bayesianism and causality, or, why I am only a half-Bayesian', in D. Corfield and J. Williamson (eds), *Foundations of Bayesianism*, Kluwer Applied Logic Series, Kluwer Academic Publishers, vol. 24, 19-36.

Pearl, J. (2011a) 'The algorithmization of counterfactuals', *Annals of Mathematics and Artificial Intelligence*, 61(1), 29-39.

Pearl, J. (2011b) 'The structural theory of causation', in Illari, Russo, and Williamson (Eds.) *Causality in the sciences*, 697-727.

Price, H. & Weslake, B (2009) 'The Time-Asymmetry of Causation', in Beebee, H., Hitchcok, C., Menzies, P. (eds), *The Oxford Handbook of Causation*.

Price, H. (1991) 'Agency and Probabilistic Causality', *British Journal for the Philosophy of Science*, **42**, pp. 157–76.

Price, H. (1992) 'Agency and Causal Asymmetry', *Mind* 101, 403: 501-520.

Price, H. (2007) 'Causal Perspectivalism', in H. Price & R. Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*, Oxford University Press.

Price, H. (2014) 'Causation, Intervention and Agency: Woodward on Menzies and Price', [Preprint], forthcoming in H. Beebee, C. Hitchcock, H. Price (eds.), *Making a Difference*, Oxford University Press.

Psillos, S. (2003) A glimpse of the secret connexion: harmonising mechanisms with counterfactuals. *Perspectives on Science*, 12(3), 288–319.

Putnam, H. (1975) *Mathematics, Matter and Method*, Cambridge: Cambridge University Press.

Rafetseder, E., Cristi-Vargas, R., Perner, J. (2010) 'Counterfactual reasoning: developing a sense of "nearest possible world"', *Child Development* 81, 1: 376-389.

Reichenbach, H. (1956) *The Direction of Time*, Berkeley and Los Angeles: University of California Press.

Reiss, J. (2009) 'Counterfactuals, Thought Experiments, and Singular Causal Inference in History', *Philosophy of Science* 76(5): 712-23.

Reiss, J. (2012a) 'Causation in the sciences: an inferentialist account' *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 43(4): 769-777, URL = <http://www.jreiss.org/papers/InferentialAccount.pdf>.

Reiss, J. (2012b) 'Counterfactuals', *Oxford Handbook of the Philosophy of Social Science* (ed. by Harold Kincaid), Oxford: OUP, 154-83.

Rips, L. (2010), ‘Two Causal Theories of Counterfactual Conditionals’, Cognitive Science 44: 2, 175-221.

Rips, L. J., & Edwards, B. (2013) ‘Inference and explanation in counterfactual reasoning’, *Cognitive Science* 37: 6.

Salmon, W. (1984) *Scientific Explanation and the Causal Structure of the World,* Princeton University Press.

Salmon, W. (1998) *Causality and Explanation*, Oxford University Press.

Schaffer, J. (2004), ‘Trumping Preemption’ Collins et al., *Causation and Counterfactuals,* The MIT Press:

Schaffer, J. (2014) ‘The Metaphysics of Causation’, *The Stanford Encyclopedia of Philosophy* (Summer 2014 Edition), Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/sum2014/entries/causation-metaphysics/>.

Schlottmann, A. (1999) ‘Seeing it happen and knowing how it works: How children understand the relation between perceptual causality and underlying mechanism’, *Developmental Psychology, 35*(5): 303–317.

Schulz, L. & Gopnik, A. (2004) ‘Causal learning across domains’, *Developmental Psychology* 40: 162-176.

Schulz, L., Bonawitz, E., Griffiths, T. (2007) ‘Can being scared cause tummy aches? Naïve theories, ambiguous evidence and preschoolers’ causal inferences’, *Developmental Psychology* 43: 1124-1139.

Schulz, L., Gopnik, A., Glymour, C. (2007) ‘Preschool children learn about causal structure from conditional interventions’, *Developmental Science* 10: 322-332.

Schwartz D.L., Black T. (1999) ‘Inferences through imagined actions: knowing by simulated doing’, *Journal of Experimental Psychology*: Learning Memory Cognition 25:116–36

Shultz, T. R. (1982) ‘Rules for causal attribution’, *Monographs of the Society for Research in Child Development,* 47 (1, Serial No. 194).

Sloman, S. & Lagando, D. (2015) ‘Causality in Thought’, *Annual Review of Psychology* 66: 3.1-3.25.

Smith, N.J.J. (2013) ‘Time Travel’, *The Stanford Encyclopedia of Philosophy* (Winter 2013 Edition), E.N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/win2013/entries/time-travel/>.

Spirtes, P., Glymour, C., & Scheines, R. (1993) *Causation Prediction and Search,* New York, NY: Springer-Verlag.

Strevens, M. (2003) ‘Against Lewis’s New Theory of Causation’, *Pacific Philosophical Quarterly* 84, 398–412.

Strevens, M. (2007) Essay review of Woodward, *Making Things Happen*, *Philosophy and Phenomenological Research* 74: 233-249.

Strevens, M. (2008a) Comments of Woodward, *Making Things Happen*, *Philosophy and Phenomenological* 77: 171-192.

Strevens, M. (2008b) *Depth: An Account of Scientific Explanation*, Harvard University Press.

Strevens, M. (2013) ‘Causality Reunified’, *Erkenntnis* 7: 299-320.

Tversky A., Kahneman D. (1974) ‘Judgment under uncertainty: heuristics and biases’, *Science* 185:1124–31.

Waskan, J. (2011) ‘Mechanistic explanation at the limit’, *Synthese* 183: 389-408.

Williamson, J. (2005) *Bayesian Nets and Causality*, Oxford University Press.

Woodward, J. & Hitchcock, (2003) ‘Explanatory Generalizations I’, in *Nous* 37:1 (2003): 1-24.

Woodward, J. (2003) *Making Things Happen*, Oxford University Press.

Woodward, J. (2006) ‘Sensitive and Insensitive Causation’, *The Philosophical Review* 115: 1 (2006): 1-50.

Woodward, J. (2007a) ‘Interventionist theories of causation in psychological perspective’, in Gopnik, A., Schulz, L. (eds), *Causal Learning*, Oxford University Press.

Woodward, J. (2007b) ‘Causation with a Human Face’, in Huw Price & Richard Corry (eds.), *Causation, Physics, and the Constitution of Reality: Russell's Republic Revisited*. Oxford University Press.

Woodward, J. (2008a) ‘Causation and Manipulability’, *The Stanford Encyclopedia of Philosophy (Winter 2008 Edition),* Edward N. Zalta (ed.), plato.stanford.edu/entries/causation-mani/ .

Woodward, J. (2008b) ‘Response to Strevens’, *Philosophy and Phenomenological Research* 77:1 (2008): 193-212.

Woodward, J. (2011a) ‘A Philosopher Looks at Tool Use and Causal Understanding’, in McCormack et al. (eds), *Tool Use and Causal Cognition*, Oxford University Press, 18-51., URL = < http://philsci-archive.pitt.edu/8653/>.

Woodward, J. (2011b) ‘Mechanisms Revisited’, *Synthese* 183: 409-427.

Woodward, J. (2012) ‘Causation: Interactions between Philosophical Theories and Psychological Research’, *Philosophy of Science* 79:5, 961-972.

Woodward, J. (2014) ‘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: 5, 691-713.