UNIVERSITY OF CALIFORNIA, SAN DIEGO

Causation and Experimentation from Philosophy to Evidence-Based Policy

A Dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Philosophy

by

Alexandre Marcellesi

Committee in charge:

Nancy Cartwright, Chair
Kate Antonovics
Craig Callender
Stephan Haggard
Michael Hardimon
Kerry McKenzie

2016

The Dissertation of Alexandre Marcellesi is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2016

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I thank my committee members, past and present, for their help and support in completing this dissertation: Kate Antonovics, Nancy Cartwright, Craig Callender, Stephan Haggard, Michael Hardimon, Kerry McKenzie, Karthik Muralidharan and Christian Wüthrich. I also thank fellow graduate students and former colleagues who, at some point, read parts of this dissertation, gave me feedback or helped me think through the ideas it contains: Craig Agule, Gil Hersch, Joyce Havstad, Casey McCoy, Chris Pariso, Ben Sheredos and Jacob Stegenga. I thank my family for supporting me and waiting until my seventh year to start pressing me about when I would finally graduate: Dahbia, Alain and Aurelia Marcellesi. Finally, I thank Katie Simpson for her unwavering emotional support.

My research has been supported, directly or indirectly, by the following institutions: The John Templeton Foundation, the UCSD Academic Senate and the UK's Arts and Humanities Research Council.

Chapters 6 and 7 have both been published in *Philosophy of Science* and are reprinted here with permission from the publisher:

- Marcellesi, A. 2013. "Is Race a Cause?" *Philosophy of Science* 80(5): 650-659.
- Marcellesi, A. 2015. "External Validity: Is There Still a Problem?" *Philosophy of Science* 82(5): 1308-1317.

Chapter 8 was coauthored with Nancy Cartwright and is included in this dissertation with her permission:

- Marcellesi, A. and Cartwright, N. 2013. "Modeling Climate Mitigation and Adaptation Policies to Predict their Effectiveness: The Limits of Randomized Controlled Trials."

VITA

2006	B.A. in Philosophy, Université Paris-Sorbonne
2009	M.A. in History and Philosophy of Science, Université Paris-Sorbonne
2009	ENS Diploma in Philosophy, École Normale Supérieure
2016	Ph.D. in Philosophy, University of California, San Diego

ABSTRACT OF THE DISSERTATION

Causation and Experimentation from Philosophy to Evidence-Based Policy

by

Alexandre Marcellesi

Doctor of Philosophy in Philosophy

University of California, San Diego, 2016

Nancy Cartwright, Chair

The view that there exists a privileged methodological relationship between experimentation and causation is widely held among scientists and philosophers alike. But some philosophers, starting from the intuition that causes 'make a difference' to their effects, think that this relationship runs deeper and is also metaphysical and semantic in nature. The question that constitutes the guiding thread running through the papers that make up this dissertation is that of the strength of the connection—metaphysical, semantic and methodological—between causation and experimentation. Just how closely are experimentation and causation related? The view I defend in this dissertation, sometimes explicitly but often implicitly, is that this connection is not nearly as tight as it is often taken to be.

Chapter 1

General Introduction

Why write a dissertation on causation? Because causation is everywhere. It is everywhere in philosophy "because of the myriad ways in which [it] works its way into a host of [...] contemporary philosophical debates: debates concerning (to name a few) mental causation and the nature of mind, epistemology, perception, color, action theory, decision theory, semantics, scientific explanation, the asymmetry of the temporal arrow, and moral and legal responsibility" (Paul and Hall, 2013, 1) But causation is also everywhere in science. As Phyllis Illari and Federica Russo put it, "Many of the sciences spend much of their time looking for causes—from causes of disease, to climate change, to earthquakes, to house price bubbles." (Illari and Russo, 2014, 3) This is especially true in the social sciences producing the knowledge that serves as the basis for policy-making. Finally, causation is everywhere in our everyday lives. Not a day goes by without causal knowledge, explicitly or implicitly, informing our decisions by helping us predict the effects of our actions. Causation thus is a central notion. We need a solid understanding of it.

What exactly do we ask when we ask 'What is causation?' Illari and Russo (2014, 238-240) distinguish five interpretations of this question:

1. *Epistemology*: What is causal knowledge and by what channels does one acquire it?

2. *Metaphysics*: What is causation 'in the world', as it were?

3. *Semantics*: What do we mean when we say that 'A causes B'?

4. *Methods*: By what methods can we discover causal relations?

5. *Use*: What is causal knowledge good for and how should we use it?

There is wide disagreement among philosophers on 1, 2 and 3. There is neither agreement nor disagreement on 5, since it is a question rarely considered. But there is a fairly broad consensus on 4, both among philosophers and among scientists who occupy themselves with testing causal hypotheses. The consensus view is that, when it comes to discovering causal relations and distinguishing them from mere correlations, experimental methods are the way to go. This has been the case at least since Francis Bacon (1878[1620]) and John Stuart Mill (1843). Today, a specific kind of experimental method, the randomized controlled trial (RCT), is widely touted as the 'gold standard' for causal inference (see e.g. Rubin 2008).

Some philosophers have argued that the connection between causation and experimentation is more than simply methodological. This is for instance the case of James Woodward who, taking his inspiration from agency theorists such as Collingwood (1940), Gasking (1955), von Wright (1971) or Menzies and Price (1993), argues that the connection between experimentation and causation is also metaphysical and semantic.

There is a controversy (see Strevens 2007, 2008a; Woodward 2008) regarding how to interpret the view Woodward develops in his magnum opus, *Making Things Happen* (2003b). It is not immediately clear which of Illari and Russo's five questions Woodward's account of causation, variously called 'manipulationist' or 'interventionist', is designed to answer. But one can confidently say that Woodward endorses the two following theses:

- Causal facts are just facts regarding the outcomes of counterfactual experimental manipulations.

- Causal claims are just claims about the outcomes of counterfactual experimental manipulations.

Woodward's view regarding the metaphysical and semantic connection between experimentation and causation is shared by prominent statisticians such as Donald Rubin (1986) and Paul Holland (1986).

This dissertation is made up of papers that were written at different times and for different purposes. The question that constitutes the guiding thread running through these papers is that of the strength of the connection—metaphysical, semantic and methodological—between causation and experimentation. Just how closely are experimentation and causation related? As will become clear in the following chapters, my view is that this connection is not nearly as tight as it is often taken to be. But before giving a brief summary of each chapter of the dissertation, let me first say a little bit more about Woodward's interventionist view, since it played an important role in the genesis of this work.

Woodward develops both an account of causation and an account of causal explanation. These accounts both revolve around the notion of an *intervention*, a notion Woodward uses to define various causal notions. Though I will characterize the notion of an intervention in more detail in Chapters 3, 4 and 5, let me here give the basic idea. An intervention on A with respect to B—where A and B might be events, facts, variables, etc. (I want to remain neutral regarding the nature of causal relata for now)—is a manipulation that alters A in some way (by removing it, introducing it, changing its size, etc.) without directly altering B or any of its causes except for those that are also effects of A.

What kinds of things are interventions? The manipulations Woodward refers to are meant to be further events, like flipping a light switch or swinging a hammer. Indeed, since an intervention on A with respect to B is a manipulation that is required to *cause* a change in A, a manipulation must be some further *physical* event. Woodward's notion of intervention thus differs from Judea Pearl's '*do*-operator' (Pearl, 2000, 70), and this

even though Woodward and Pearl are routinely, but misleadingly, lumped together under the 'interventionist' label (see e.g. Campbell 2007, 58, Kuorikoski 2014, 334 or Menzies 2012, 800). To put it briefly, whereas interventions are physical events on Woodward's view, applications of the *do*-operator are, for Pearl, transformations of formal models representing systems of causal relations. Though these two notions are related, they differ in crucial ways, as Woodward (2003a, 110) himself notes.

Woodward is also often lumped together with the Carnegie Mellon trio of Peter Spirtes, Clark Glymour and Richard Scheines because these authors use a notion of intervention similar to Woodward's in (Spirtes et al., 2001). Again, this grouping is misleading since, as Glymour notes in his review of (Woodward, 2003b), neither he nor his colleagues, unlike Woodward, purport to use the notion of intervention to *define* causation (Glymour, 2004, 790).

How then does Woodward use the notion of intervention to define causation? Again, I will here limit myself to an informal presentation of Woodward's views, since they will be described in more detail in the following chapters. The basic idea is simple: If an intervention on A with respect to B occurs and is followed (temporally) by a change of any kind in B, then the change in A must be causally responsible for this change in B, and so A must be a cause of B. Since an intervention on A with respect to B will change A and nothing else, then, provided one rules out the possibility of spontaneous and uncaused changes in B, the culprit for the change in B cannot but be A. Underlying Woodward's account of causation is the intuition, shared by counterfactual accounts of the kind pioneered by David Lewis (1973a) and probabilistic accounts of kind pioneered by Patrick Suppes (1970), that causes 'make a difference' to their effects: When causes change, so do their effects.

The initial aim of this dissertation was to critically examine Woodward's account of causation, i.e. the set of conditions he advances as being both necessary and sufficient for

a genuine causal relation to exist. As originally stated, Woodward's account of causation faces a number of issues, detailed for instance by Michael Strevens (2007; 2008a), Michael Baumgartner (2009; 2012) or Alexander Reutlinger (2013).

Though my intention was to pile on and add my voice to the (small) chorus of critics of Woodward's account of causation, I ended up changing my aim because, as Nancy Cartwright (2007, Chapter 10) or Reutlinger (2013) have argued, Woodward's interventionist account is little more than a conceptual variant of already existing probabilistic and counterfactual account. It recycles concepts, centrally the notion of intervention developed by Spirtes, Glymour and Scheines, to formulate definitions that are, in substance, equivalent to definitions already on offer or to simpler definitions that do not require an appeal to the notion of intervention.

Rather than focusing on Woodward's account of causation, then, I decided to switch my focus to his account of causal explanation and, more generally, to broaden the scope of my dissertation. Let me tell you how I did so by giving you a brief summary of each chapter.

Chapter 2 *David Lewis: Causality and Counterfactuals*

As I noted above, the intuition that causes 'make a difference' to their effects is at the root of the accounts of causation defended by both Woodward and Lewis. Lewis uses the technical apparatus of possible worlds and his own analysis of counterfactual conditionals in order to develop this intuition into a full-blown account of causation.

This chapter has multiple aims. The first is to introduce the reader to Lewis' analysis of counterfactuals and to his two definitions of causation (from 1973 and from 2000). The second is to examine the limitations of Lewis' account. The third is to look at more recent accounts of causation that are set in the structural equations framework and are supposed to supersede Lewis' more traditional counterfactual account. And the fourth is to argue that such accounts face serious difficulties that are both conceptual and methodological. I focus on an account advanced by Christopher Hitchcock (2007) to develop these arguments.

The lesson of Chapter 2 is that, it seems, causation involves more than just causes 'making a difference' to their effects. Contrary to what Lewis hoped, a reduction of causation to counterfactual dependence looks unlikely and the methodology adopted by Lewis and his disciples was unlikely to help us achieve such a reduction in the first place.

Chapter 3 *Interventions, Invariance and Explanatory Relevance: Not So Fast*

The interventionist account of causal explanation developed by Woodward, in collaboration with Hitchcock (Woodward, 2003b; Woodward and Hitchcock, 2003; Hitchcock and Woodward, 2003), is one of the most popular and influential on offer. In Chapter 3 I argue that, despite claims to the contrary, it does not solve Wesley Salmon's problem of explanatory relevance, a problem it was explicitly designed to solve. Because it fails to solve the problem of explanatory relevance, the interventionist account fails to properly draw the line between information that is causally explanatory and information that is not. It is therefore inadequate as an account of causal explanation. I also show that, by contrast with the interventionist account, Strevens's kairetic account (2008b) does solve the problem of explanatory relevance provided it is tweaked in one minor way.

The lesson of Chapter 3 is that to causally explain an event one must do more than just describe the outcomes of counterfactual experimental manipulations.

Chapter 4 *Probing the Depths of Explanatory Depth*

Not all causal explanations are created equal and some are better than others. But what are the factors the quality of a causal explanation depends on? It is to answer this question that Woodward and Hitchcock have developed an account of what they call 'explanatory depth'. David Harker (2012) has suggested that depth, as Woodward and Hitchcock define it, is a "peculiarly explanatory virtue" advocates of Inference to the Best Explanation can appeal to in order to flesh out their account.

In Chapter 4 I argue that the notion Woodward and Hitchcock provide an account of

is just predictive power by another name and is not properly seen as an explanatory notion. As a result, it is not a notion advocates of Inference to the Best Explanation should appeal to. I also argue that Woodward and Hitchcock's account conflicts with the view that causal explanations are better when they cite causes that are 'proportional', in Stephen Yablo's sense, to their effects, a view many—including Woodward—take to be plausible. If the arguments I develop are sound, then there are good reasons to think that the account of explanatory depth developed by Woodward and Hitchcock is inadequate.

The lesson of Chapter 4 is that causal explanations are not better when they better describe the outcomes of counterfactual experimental manipulations.

Chapter 5 *Interventionism Does Not Explain the Practical Usefulness of Causal Knowledge*

Woodward defends the view that causal knowledge is practically useful because it helps us predict the effects of our actions and thus guides our decision-making. This view is not original to Woodward: It has been the received view in the contemporary philosophical literature at least since Nancy Cartwright's seminal "Causal Laws and Effective Strategies" (1979).

But Woodward also claims that his popular and influential interventionist account of causation provides an explanation for the practical usefulness of causal knowledge thus understood, a bridge leading from causal claims to predictions regarding the effects of our actions. And, indeed, providing such an explanation is often taken to be one of the main qualities of the interventionist account.

In Chapter 5 I argue that, despite appearances—and claims—to the contrary, Woodward's interventionist account of causation does not explain the practical usefulness of causal knowledge and, in particular, its relevance to decision-making, and this even though providing such an explanation is widely assumed to be one of its main qualities.

The lesson of Chapter 5 is that identifying causal facts with facts regarding the out-

comes of counterfactual experimental manipulations is not enough to explain the relevance of causal knowledge to our actual manipulations.

Chapter 6 *Is Race a Cause?*

Advocates of the counterfactual approach to causal inference such as Rubin and Holland follow Woodward in tying causal claims to claims regarding the outcomes of counterfactual experimental manipulations. And, on the basis of this view, they have argued that race is not a cause, and this despite the fact that it is commonly treated as such by scientists in many disciplines.

In Chapter 6 I object that the argument developed by Rubin, Holland and others is unsound since two of its premises are false. I also sketch an argument to the effect that racial discrimination cannot be explained unless one assumes race to be a cause.

The lesson of Chapter 6 is that drawing too tight a conceptual connection between causation and experimentation can lead one astray, as it does with Holland and Rubin in the case of race, and so that the view of causation one adopts can have serious concrete ramifications.

Chapter 7 *External Validity: Is There Still a Problem?*

Chapter 7 and 8 are concerned with the methodological aspect of the connection between causation and experimentation, and my focus therefore turns away from philosophy and toward evidence-based policy. But there is a straightforward connection between philosophical views and methodological principles. If causal hypotheses are just claims regarding the outcomes of hypothetical experiments, then what better way to test them than to actually carry out these experiments? It thus seems that the view of causation advocated by Woodward and others can provide a foundation for the claim that controlled experiments, and RCTs in particular, are the gold standard for causal inference. Chapter 7 questions this gold standard claim implicitly whereas Chapter 8 does so explicitly.

The topic of Chapter 7 is the problem of external validity as it arises in evidence-based policy. I first propose to distinguish between two kinds of external validity inferences, predictive and explanatory. I then argue that we have a satisfactory answer to the question of the conditions under which predictive external validity inferences are good inferences. If this claim is correct, then it has two immediate consequences: First, some external validity inferences are deductive, contrary to what is commonly assumed (see e.g. Guala 2005, 196). Second, Daniel Steel's requirement that an account of external validity inference break what he calls the 'Extrapolator's Circle' (Steel, 2008, 4) is misplaced, at least when it comes to predictive external validity inferences.

There is a third consequence, and it is the lesson from Chapter 7: If one's aim in testing a hypothesis regarding the effects of some policy is to inform subsequent policy decisions, as is often the case, then one needs more than just the knowledge typically produced by running an RCT.

Chapter 8 *Modeling Mitigation and Adaptation Policies to Predict their Effectiveness: The Limits of Randomized Controlled Trials (with Nancy Cartwright)*

The final chapter of this dissertation, Chapter 8, was written in collaboration with Nancy Cartwright. In this chapter we describe the logic behind RCTs and explain why they are widely assumed to be the gold standard for causal inference. But we argue that, when it comes to evaluating the effects of climate change policies, be they adaptation policies or mitigation policies, RCTs face serious difficulties. We illustrate this claim by reviewing a number of climate policies, focusing in particular on Payment for Environmental Services policies that are currently popular. We also show that, in some cases, the imperative to measure the effects of policies with accuracy can be counterproductive and undermine the policies in question.

The lesson of Chapter 8 is that one needs more tools than just experiments in order to evaluate the effects of climate policies and so that calls for more experimental evaluations

of these policies, at the exclusion of other methodologies, are misguided.

Chapter 2

David Lewis: Causality and Counterfactuals

2.1 Introduction

As you may remember, Lehman Brothers filed for bankruptcy in September 2008. The subprime mortgage crisis that occurred in the US in 2007-2008 is one of the many causes of this bankruptcy. But what do we mean when we call the former event a *cause* of the latter? And what is the nature of the relation that, we are claiming, exists between them? One the projects David Lewis was engaged in from 1973 and right up until his death in 2001 was to provide an analysis of our concept of causation. We routinely formulate causal judgments. This is what I did above in relating Lehman Brothers' bankruptcy to the subprime mortgage crisis. Forming such judgments requires us to apply our concept of causation. One way to learn about the nature of this concept is to formulate a definition of it, i.e. a list of conditions that are necessary and sufficient for its application to be appropriate, and to check the adequacy of this definition by confronting it with our causal judgments. If this definition conflicts with too many of our causal judgments—in the sense that it yields the verdict that A causes B when our intuitions tell us otherwise—than it is discarded and

replaced by a new and, hopefully, improved definition that is once again put to the test. Once we arrive, through this process, at a definition that agrees with the vast majority of our causal judgments, we will have successfully identified the boundaries of our concept of causation. This is, in a nutshell, the methodology Lewis adopts to carry out the project of a conceptual analysis of causation.

Conceptual analysis, however, is not the end goal for Lewis but, rather, a stepping stone on the way to a metaphysical reduction of causal facts. If one can formulate a definition of causation that is both adequate and reductive, in the sense that its definiens is free of causal notions, then one is in a good position to argue that the relation picked out by our concept of causation is not primitive and that it can be broken down into components that are not themselves causal.[1]

The aim of this paper is largely expository. I will give a brief and incomplete overview of Lewis' attempts to define causation. I will also describe, and briefly criticize, a more recent attempt due to Christopher Hitchcock (2007). As I will explain, Lewis and Hitchcock are engaged in distinct projects, and this even though they employ very much the same methodology. Did Lewis succeed in completing his project of a conceptual analysis and metaphysical reduction of causation? No, for reasons I will detail below. Will this project ever be successful in the future? I will, on the basis of arguments developed in (Glymour et al., 2010), defend the view that the methodology employed by Lewis and his successors makes such a success unlikely. As I will explain, the same worry arises for philosophers, e.g. Hitchcock, who adopt Lewis' methodology in order to carry out distinct projects.

Let me make four remarks before diving into the details of Lewis' views about causation. First, Lewis was interested in causation understood as a relation holding between actual events, where events are assumed to be particulars located in space and time. This

[1] See, e.g., (Nolan, 2005, chapter 5) for a much more detailed and precise description of Lewis' methodology.

relation is now often called ‘singular causation’, ‘token causation’ or, more and more frequently, ‘actual causation’, to distinguish it from other causal relations.[2] When I use the word ‘causation’ below, therefore, it will be to refer to this specific causal relation. Second, I will follow Lewis in primarily focusing my attention on deterministic cases, i.e. on cases in which causes guarantee the occurrence of their effects and do not simply alter the probability that they will occur. Third, I will say nothing below about issues such as the transitivity of causation or the causal status of absences or omissions. Fourth, I will here ignore objections—such as the ones advanced in (Dowe, 2000, chapter 1) or (Woodward, 2014)—that target not Lewis’ definitions of causation but the very idea that providing a conceptual analysis of causation is a worthy endeavor.

What follows is not intended to be an exhaustive summary of Lewis’ views on causation, of the objections they face or of the answers that have been formulated by Lewis and others. The size of the literature on this topic is overwhelming and, in any case, much too large to be reviewed in a single chapter. I direct the interested reader to (Nolan, 2005, chapter 4) and especially (Menzies, 2014) as good entry points into this literature.

2.2 The 1973 analysis

What do we mean when we say that some event c is a cause of another event e? How should one understand the concept of causation involve in such a judgment? Lewis’ views about causation are premised on the intuition expressed in the following quotation: “We think of a cause as something that makes a difference, and the difference it makes must be a difference from what would have happened without it.” (Lewis, 1973a, 557) What we mean when we say, for instance, that the subprime mortgage crisis is a cause of the bankruptcy of Lehman Brothers is that the occurrence of the former event ‘made a difference’ to the

[2] See for instance the recent special issue of *Erkenntnis* on ‘Actual Causation’ (2013, volume 78, issue 1 supplement).

occurrence of the latter. Had the subprime mortgage crisis not occurred, we are inclined to think, Lehman Brothers would not have gone bankrupt. But several steps separate this basic intuition from a definition of causation. Below I detail the steps involved in Lewis' first attempt at a definition of causation, developed in (Lewis, 1973a).

Step 1 Since Lewis' definition of causation appeal to counterfactual conditionals (or, more briefly, counterfactuals), the first step in constructing such a definition is to provide truth-conditions for such propositions.[3] Take any two propositions A and B and call the proposition that if A were the case, then B would be the case their *counterfactual*. Under what conditions is this proposition—in symbols, $A \mathrel{\square\!\!\rightarrow} B$—true? Lewis takes the intuition that counterfactuals are about unactualized possibilities at face value and appeals to his machinery of possible worlds (see e.g. Lewis 1986b, Chapter 1) in order to specify their truth-conditions.[4]

There are two ways for a counterfactual $A \mathrel{\square\!\!\rightarrow} B$ to be true at a world w. Call a world in which a proposition A is true an A-world. The first way for $A \mathrel{\square\!\!\rightarrow} B$ to be true at w is for it to have an impossible antecedent, i.e. for there to be no possible A-world relative to w.[5] If it is impossible for Nicolas Sarkozy to be a sheep, then, according to Lewis' analysis of counterfactuals, it is (vacuously) true that if Sarkozy were to be a sheep, he would produce 15 pounds of wool per year. Note that whenever, in what follows, I do not explicitly specify the world relative to which the truth-value of some counterfactual is to be evaluated, I will assume—as I just did with the counterfactual regarding Sarkozy's wool production—that

[3] It should be noted that, for Lewis, propositions are not linguistic entities. On his view, the proposition that p is identical with the set of possible worlds at which it is true that p. The proposition that snow is white, for instance, is identified with the set of possible worlds at which snow is white. The English sentence 'Snow is white' *expresses* the proposition that snow is white because the former is true at exactly the same worlds as the latter.

[4] I will here ignore debates regarding the ontological status of possible worlds. See (Divers, 2002) for a detailed introduction.

[5] For Lewis, the notion of possibility is, like the notion of truth, relativized to worlds. It is therefore possible for the antecedent of a counterfactual to be impossible relative to some world w_1 and possible relative to some other world w_2. I will leave the relativization of the notion possibility implicit in what follows. Note also that the relevant notion of possibility for Lewis here is that of *metaphysical* possibility.

the world in question is the actual world (or @).

What when there are possible A-worlds and $A \boxdot\!\!\rightarrow B$ therefore has a possible antecedent? Call a world at which both A and B are true an $A \wedge B$-world. And call a world at which B is false a $\neg B$-world. A counterfactual $A \boxdot\!\!\rightarrow B$ with a possible antecedent is true at a world w if and only if (or iff) all of the A-worlds closest to w also are B worlds.[6] Or, to put it slightly differently, $A \boxdot\!\!\rightarrow B$ is true at w iff some $A \wedge B$-world is closer to w than is any $A \wedge \neg B$-world. What does Lewis mean when he talks about closeness between possible worlds? On his view, w_1 is closer to w than is w_2 iff w_1 is, all things considered, more similar to w than w_2 is.

Though Lewis' official position is to take the relation of overall comparative similarity between worlds as a primitive, he nonetheless singles out certain respects of similarity (e.g. similarities regarding laws of nature and particular facts) as especially important for evaluating the counterfactuals that are to be used in analyzing causation (see Lewis 1973a, 560; 1979, 472). I will come back to this point below, for it will become important when we examine the way Lewis proposes to deal with putative counterexamples to his definitions of causation. Let me here say just one more thing about the relation of overall comparative similarity: Lewis assumes that, for any world w, the world closest to w is w itself, and this regardless of the counterfactual antecedent one is considering. Though this assumption has been criticized (see e.g. List and Menzies 2009, 484–486), it is intuitively compelling: How could any world w ($w \neq @$) be as similar (a fortiori, more similar) to @, for instance, as @ itself is?

Consider, as an illustration of Lewis' account of the truth-conditions of counterfactuals, the proposition that if Sarkozy were 7 feet tall, then he would not wear stacked heels. Assuming that there is some possible world in which Sarkozy is 7 feet tall, under which conditions will this counterfactual be true? According to Lewis' account, this counterfactual

[6] There need not be a single closest A-world to w. Lewis (1973b, 20–21) rejects what is commonly known as the *Limit Assumption*.

is true iff all of the closest worlds at which Sarkozy is 7 feet tall are worlds at which he does not wear stacked heels. In other words, this counterfactual is true iff some world at which Sarkozy is 7 feet tall and does not wear stacked heels is overall more similar to @ than any world at which Sarkozy is 7 feet tall but nonetheless wears stacked heels. Now that Lewis' account of counterfactuals has been briefly introduced and illustrated, we can move on to the next step toward his definition of causation.

Step 2 Call two propositions A and B *compossible* iff it is possible for both to be true at the same world. A family of propositions $B_1, B_2, \ldots, B_n$, no two of which are compossible, *counterfactually depends* upon another such family $A_1, A_2, \ldots, A_n$ iff $A_i \mathrel{\Box\!\!\rightarrow} B_i$ for every $i = 1, \ldots, n$.

As I indicated above, Lewis is after a definition of causation understood as a relation holding between events. All I have talked about so far, however, are propositions. Fortunately, it is easy to establish a connection between events and propositions. As Lewis (1973a, 562) puts it, "To any possible event e, there corresponds the proposition $O(e)$ that holds at all and only those worlds where e occurs. This $O(e)$ is the proposition that e occurs." Given this connection, one can easily extend the definition of counterfactual dependence given above from propositions to events: A family of events $e_1^*, e_2^*, \ldots, e_n^*$ (subject to the compossibility restriction) counterfactually depends on another such family of events $e_1, e_2, \ldots, e_n$ iff the family of propositions $O(e_1^*), O(e_2^*), \ldots, O(e_n^*)$ counterfactually depends on the family of propositions $O(e_1), O(e_2), \ldots, O(e_n)$. What does counterfactual dependence between events thus defined have to do with causation? The identification of counterfactual dependence with causal dependence is the next step in the construction of Lewis' definition.

Step 3 As Lewis (1973a, 561) notes, counterfactual dependence "between large families of alternatives is characteristic of processes of measurement, perception, or control." And we typically think of these processes as causal, even when the families in question are rather

small. Consider the limit case in which each family of events has just two members, with the c's corresponding to the state of the light switch in some room ($c_1 =$ up, $c_2 =$ down) and the e's corresponding to the state of the light in the same room ($e_1 =$ on, $e_2 =$ off). The fact that both $c_1 \mathbin{\Box\!\!\rightarrow} e_1$ and $c_2 \mathbin{\Box\!\!\rightarrow} e_2$ are true, let me assume, certainly seems to incline us to think that the state of the switch is causally related to the state of the light.

This intuition is stronger the larger the families of events standing in a relation of counterfactual dependence. Readings of a functioning thermometer, for instance, will counterfactually depend upon the temperature of the medium surrounding the thermometer. And the fact that they do so, one is inclined to infer, is indicative of a causal relation between the temperature of the medium and the state of the thermometer. It is on the basis of the idea that 'concomitant variations', to use John Stuart Mill's term, across a range of counterfactual circumstances indicates the presence of a causal relation that Lewis proposes to identify counterfactual dependence between families of events with *causal dependence*. The next step on the way to Lewis' definition of causation is to define causal dependence between single events rather than between families of events.

Step 4 Take two events c and e and call $\neg O(e)$ the proposition that event e does not occur. According to Lewis, e causally depends on c iff the family $O(e), \neg O(e)$ counterfactually depends on the family $O(c), \neg O(c)$, i.e. iff the following two counterfactuals are true:

(i) $O(c) \mathbin{\Box\!\!\rightarrow} O(e)$

(ii) $\neg O(c) \mathbin{\Box\!\!\rightarrow} \neg O(e)$

Consider the case in which c and e actually occur and so in which $O(c)$ and $O(e)$ are actually true. In this case, (i) is automatically true: Since no world is as close to @ as @ itself is, the closest $O(c)$-world is @ itself, and this world also is an $O(e)$-world. In other words, whenever two propositions A and B are true at a world w, the counterfactuals $A \mathbin{\Box\!\!\rightarrow} B$ and

$B \mathrel{\Box\!\!\rightarrow} A$ are also true at w.[7]

Consider again the example of the subprime mortgage crisis and the bankruptcy of Lehman Brothers. Both of these events actually occurred, unfortunately. According to Lewis' account of counterfactuals, this implies the truth of the proposition that if the subprime mortgage crisis were to happen, then Lehman Brothers would go bankrupt. If it is also a true that had the subprime mortgage crisis not occurred, Lehman Brothers would not have gone bankrupt, then the latter event causally depends on the former. But causal dependence between single events is not yet causation. Note, for instance, that two events neither of which actually occurs can stand in a relation of causal dependence. We do not, however, normally think of non-actual events, e.g. the event of Ségolène Royal being elected President of France in 2007, as being causes. The next—and last—step will take us from causal dependence to causation.

Step 5 According to Lewis, causal dependence between actual events implies causation. If c and e actually occur and e causally depends on c, then c is a cause of e. If it is true that had the subprime mortgage crisis not occurred, Lehman Brothers would not have gone bankrupt then the former event is a cause of the latter. Note that I say *a* cause and not *the* cause. Lehman Brothers' bankruptcy had other causes besides the subprime mortgage crisis, e.g. a series of bad decisions on the part of its board of directors. As far as Lewis' definition of causation is concerned, there is no sense in asking which of these was *the* cause of the bankruptcy and which ones were mere background conditions for it. All of these events are, with respect to causation, on a par. Lewis thus has nothing to say about what he calls "principles of invidious discrimination" (1973a, 595).

Are we home free yet, now that causation has been identified with causal dependence among actual events? No, but we are almost there. Lewis assumes that causation is transitive:

[7]Note that this implies that, despite their name, counterfactuals need not have false antecedents, i.e. they need not have antecedents that run 'counter to the facts'.

If c causes d and d causes e, then c causes e (see e.g. Hitchcock 2001 for objections to this view). But causal dependence as defined above is not transitive: It is possible for e to causally depend on d and d to causally depend on c without e causally depending on c.[8] There is an easy way around this difficulty, however. Call a finite sequence of events e_1, e_2, e_3, ..., e_n a *causal chain* just in case e_2 causally depends on e_1, e_3 causally depends on e_2, and so on all the way up to e_n. Lewis simply defines *causation* as follows: An actual event c is a cause of another actual event e iff there is a causal chain leading from the former to the latter. Causation is what Lewis (2000, 184) calls the *ancestral* of the relation of causal dependence.[9]

How is the definition of causation just introduced to be evaluated? As I indicated in Section 2.1, Lewis is after a definition of causation that is both adequate and reductive. The first step in evaluating Lewis' definition of causation is thus is check for its adequacy. Only later should one worry about its reductive character. Since what Lewis seeks to define is *our concept* of causation, one is to check for the adequacy of the definition he offers by confronting it with the intuitive judgments, or "naive opinions" (1973a, 567, n. 12), we formulate using this concept. Cases in which our intuitions conflict with Lewis' definition constitute counterexamples to this definition. And cases in which our intuitions agree with Lewis' definition are evidence in its support. What, then, is the evidence in favor Lewis' 1973 definition of causation?

Assume that c causes e and, moreover, that there is no other possible cause of e. In this case, it seems to be true that had e not occurred, neither would have c. And this, Lewis' definition of causation tells you, means that e is a cause of c. Causation, however, is widely believed to be an asymmetric relation: If c causes e, then e does not cause c. Why believe causation to be asymmetric? In part because, it is commonly assumed, causes precede their

[8] See (Lewis, 1973b, 32–35) for some counter-examples.

[9] The notion of an ancestral relation, first defined by Gottlob Frege (1879, §26), is rather intuitive: Causation is the ancestral of causal dependence because c is a cause of e iff $c$is an *ancestor* of e in a causal chain, i.e. in a chain of stepwise causal dependence.

effects in time. If c causes e, it is also temporally prior to it. And if e is temporally posterior to c, then it cannot, by assumption, be a cause of it. The case described above thus seems to be a straightforward counterexample to Lewis' definition of causation. This is what Lewis calls the *problem of effects*.

Consider another case: Assume that c is a common cause of e and f, with f occurring after e, but that neither e nor f is a cause of the other. If one also assumes that both e and f can only be caused by c, then it seems to be true that had e not occurred, neither would have f. This is because had e not occurred, c would not have occurred, which in turn implies that f would not have occurred. If f counterfactually depends on e, however, then Lewis' analysis implies that e is a cause of f, contradicting the assumption made above. This is what Lewis call the *problem of epiphenomena*.

One could solve the problem of effects by adding to Lewis' definition of causation a clause stipulating that causes must be temporally prior to their effects. Lewis rejects this solution, however, for several reasons. First, it would rule out a priori the possibility of cases of backward (in time) causation, i.e. of cases in which effects temporally precede their causes. Second, it would make it impossible to analyze the direction of time in terms of the direction of causation. Third, such a solution would not help solve the problem of epiphenomena, since in the case described above e is temporally prior to f. How then does Lewis purport to solve the problems of effects and epiphenomena faced by his definition of causation?

The solution Lewis advocates is straightforward. As he puts it, "The proper solution to both problems, I think, is flatly to deny the counterfactuals that cause the trouble." (1973a, 566) Consider again the problem of effects: I said above that, given the assumptions made regarding c and e, it seems true that had e not occurred, neither would have c. Lewis simply denies that this is the case. But on what grounds? Remember that the truth-values of counterfactuals depend on the relevant relation of overall comparative similarity between

worlds. Since, we have assumed, c and e both actually occur, c causally depends on e just in case the closest $\neg e$-world is a $\neg c$-world. But this is precisely what Lewis denies. In his own words,

> If e had been absent, it is not that c would have been absent.... Rather, c would have occurred just as it did but would have failed to cause e. It is less of a departure from actuality to get rid of e by holding c fixed and giving up some or other of the laws and circumstances in virtue of which c could not have failed to cause e, rather than to hold those laws and circumstances fixed and get rid of e by going back and abolishing its cause c.

What Lewis denies in this passage is that the counterfactual $\neg e \boxdot\!\!\rightarrow \neg c$ should be given what he calls a "back-tracking" interpretation (Lewis, 1979, 457). Let me explain in more detail Lewis' argument in favor of a non-back-tracking interpretation.

On Lewis' view of laws of nature (see e.g. 1983, 365–368), deterministic laws (logically) determine what happens—which events occur when and where. A possible world in which e does not occur therefore is not a world in which the laws of nature holding in the actual world also hold. In other words, some violation of these laws—what Lewis (1973a, 560) calls a "miracle"—must be involved in moving from the actual world to such a possible world. Assume that c occurs at time t_0 and e at time t_1. And compare the two following worlds:

- In world w_1, the required miracle occurs at time $t_{0-\varepsilon}$ shortly before t_0. In this first world, neither c nor e occurs.

- In world w_2, by contrast, the required miracle occurs at time $t_{1-\varepsilon}$ shortly before t_1 and after t_0. In this second world, e does not occur but c does.

What Lewis claims is that w_2 is overall more similar to @ than is w_1. Both worlds differ from @ with respect to laws of nature, since both of their histories involve what, from the point of view of @, are miracles. But the history of w_2 matches that of @ all the way up to $t_{1-\varepsilon}$ whereas that of w_1 only matches it up to $t_{0-\varepsilon}$. If worlds such as w_2 are overall more

similar to @ than are worlds such as w_1, however, then all of the closest $\neg e$-worlds are c-worlds, and so it is not true that had e not happened, c would not have happened either. As a consequence, it is not the case that c causally depends on (a fortiori, is an effect of) e according to Lewis' definition of causation.

As I mentioned above, this solution—adopting a view of the overall comparative similarity relation that excludes a back-tracking interpretation of counterfactuals—applies to both the problem of effects and the problem of epiphenomena. How so for the latter? Here, remember, the counterfactual of interest is $\neg e \mathrel{\Box\!\!\rightarrow} \neg f$. Lewis' aim is to show that, contrary to what one might initially think given his description the case, this counterfactual is false. Assume that c occurs at time t_0, e at time t_1 and f at time t_2 where, remember, the only causal connection between e and f is via their common cause c. What Lewis argues is that a world in which the miracle wiping out e occurs at a time $t_{1-\varepsilon}$ shortly before t_1 but after t_0 is overall more similar to @ than is a world in which the required miracle occurs at some time $t_{0-\varepsilon}$ shortly before t_0. If this is true then all of the closest $\neg e$-worlds are $c \wedge f$-worlds, and so it is not true that had e not happened, f would not have happened either. Which in turn implies that f does not causally depend on (a fortiori, is not an effect of) e according to Lewis' definition.[10]

The fact that Lewis' analysis solves the problems of effects and epiphenomena counts as evidence in its favor. We know that it will deliver verdicts that are in agreement with our "naive opinions" about causation in cases that are analogous to the ones described above. But note that the worry raised by both problems was that Lewis' definition might erroneously yield the result that some event e_1 is a cause of another event e_2 when our intuitions tell us that this is not the case. This is not, of course, the only way for a definition of causation to fail. Lewis' definition will also fail if there are cases in which it yields the result that e_1 is not a cause of e_2 when our intuitions tell us otherwise. The definition offered

[10]Note that when I say 'counterfactual dependence' in what follows, I will mean 'counterfactual dependence under a non-back-tracking interpretation of counterfactuals'.

by Lewis will only succeed if causal dependence between actual events, as defined above, is both sufficient and necessary for causation.

So, are there any such cases in the offing? Yes. Cases involving what Lewis (2000, 182) calls "redundant causation" are prima facie problematic for his definition of causation. These are cases in which "two separate potential causes for a certain effect are both present; and either one by itself would have been followed by the effect; and so the effect depends upon neither." (2000, 182) Redundant causation, however, comes in many different flavors, three of which I will describe below. Consider the following case:

> Suzy throws a rock at a bottle, shattering it. Billy, who is standing nearby, would have thrown another rock at the bottle had Suzy not thrown hers, and Billy's rock, too, would have shattered the bottle.

Though our intuitions tell us that Suzy's throw is a cause of the bottle shattering, it seems, at least at first sight, that the shattering does not causally depend on Suzy's throw. It is not true that had Suzy not thrown her rock, the bottle would not have shattered. This is because had Suzy not thrown her rock, Billy would have thrown his and because, by assumption, Billy's throw would have resulted in the bottle shattering. This is what Lewis (2000, 184) calls a case of "early preemption".

Lewis' definition of causation has little trouble handling cases of early preemption. All one needs is to assume that there is some intermediate event d between Suzy's throw and the bottle shattering—e.g. the event of the rock occupying a certain region of space between where Suzy and Billy are standing and the bottle—such that the shattering causally depends on d and d causally depends on Suzy's throw. If it is true that (i) had Suzy not thrown her rock, d would not have occurred and (ii) had d not occurred, the bottle would not have shattered, then we have a causal chain leading from Suzy's throw to the bottle shattering and, causation being the ancestral of causal dependence, Suzy's throw is a cause of the bottle shattering.[11]

[11]Lewis' argument for the truth of counterfactual (ii) is analogous to the one he develops in providing a

Cases of early preemption, however, are what Lewis (2000, 184) calls "easy cases" of redundant causation. Consider the following variant of the case introduced above, due to Ned Hall (Hall, 2004, 235):

> Suzy and Billy each throw a rock at a bottle but Suzy's rock arrives first, shattering the bottle. Billy's rock arrives just an instant later and sails through the empty space previously occupied by the bottle. Had Suzy not thrown her rock, Billy's throw would have shattered the bottle.

This case is, in Lewis' terms, one of "late preemption" (2000, 184). And the strategy Lewis uses to deal with cases of early preemption cannot be used here. This is because the bottle shattering does not causally depend on any event *d* that is intermediate between Suzy's throw and the shattering itself. Given Billy's throw, the bottle would have shattered even if none of those events had occurred. There is thus no causal chain leading from Suzy's throw to the bottle shattering. As a consequence, Lewis' definition yields the verdict that the former is not a cause of the latter, contrary to what our intuitions tell us.

How, then, does Lewis propose to deal with cases of late preemption? His original proposal, developed in (Lewis, 1986c, 206–207), is as follows: Consider a merely possible world in which the same scenario as that described in the previous paragraph unfolds *except* for one important difference: In this merely possible world, call it *w*, Billy and his rock are entirely absent. Now, in *w*, there is a causal chain leading from Suzy's throw to the bottle shattering, and so the former is a cause of the latter (at *w*) according to Lewis' definition. But causation presumably is an intrinsic relation between events. Whether or not Billy throws his rock should not make a difference to whether or not Suzy's throw is a cause of the bottle shattering. Since the actual sequence of events leading from Suzy's throw to the bottle shattering is what Lewis (2000, 184) calls an "intrinsic duplicate" of the merely possible sequence of events occurring in *w*, it too must be a causal chain. Lewis thus adopts the view

---

solution to the problems of effects and epiphenomena. To put it briefly, according to Lewis, a world in which *d* does not occur but Suzy nevertheless throws her rock is overall more similar to the actual world than is a world in which *d* does not occur and Suzy does not throw her rock (such a world being one in which, by assumption, Billy throws his rock and the bottle shatters).

that sequences of events that are intrinsic duplicates of causal chains also are causal chains. These sequences of events, e.g. the one leading from Suzy's actual throw to the actual shattering of the bottle, exhibit what Lewis (2000, 184) calls "quasi-dependence": They qualify "as causal by courtesy", as he puts it. The concept of quasi-dependence thus helps Lewis deal with cases of late preemption: Because there is a chain of quasi-dependence leading from Suzy's throw to the shattering, the former is a cause of the latter, in agreement with our intuitions.

Are there other types of cases of redundant causation, besides cases of early and late preemption? Yes, for instance cases of trumping preemption first introduced by Jonathan Schaffer (2000). Here is the scenario (attributed to Bas van Fraassen) Lewis uses to illustrate trumping preemption:

> The sergeant and the major are shouting orders at the soldiers. The soldiers know that in case of conflict, they must obey the superior officer. But as it happens, there is no conflict. Sergeant and major simultaneously shout 'Advance!'; the soldiers hear them both; the soldiers advance. Their advancing is redundantly caused: if the sergeant had shouted 'Advance!' and the major had been silent, or if the major had shouted 'Advance!' and the sergeant had been silent, the soldiers would still have advanced. But the redundancy is asymmetrical: since the soldiers obey the superior officer, they advance because the major orders them to, not because the sergeant does. The major preempts the sergeant in causing them to advance. The major's order *trumps* the sergeant's. (Lewis, 2000, 183, emphasis original)

The intuition here is supposed to be that the major's order is a cause of the soldiers advancing while the sergeant's order is not. The soldiers advancing, however, does not causally depend on the major's order, since the soldiers would have followed the sergeant's order, and so would have advanced, had the major not given his order. The same is true of the sergeant's order. As a result, neither order is a cause of the soldiers advancing according to Lewis' definition.

Can one here appeal to quasi-dependence? Consider a world w_1 in which an intrinsic duplicate of the sequence of events leading from the major's order to the soldiers advancing

occurs but in which the sergeant and his order are absent. In w_1, the soldiers advancing causally depends on the major's order, and so the latter is a cause of the former (at w_1). What this means is that the actual event of the soldiers advancing quasi-depends on the actual order given by the major, which in turn implies that the major's order is a cause of the soldiers advancing. So far, so good.

But consider now a world w_2 in which an intrinsic duplicate of the sequence of events leading from the sergeant's order to the soldiers advancing occurs but in which the major and his order are absent. In w_2, the soldiers advancing causally depends on the sergeant's order, and so the latter is a cause of the former (at w_2). This means that the soldiers advancing in the actual world quasi-depends on the sergeant's actual order and, as a consequence, that the sergeant's order is a cause of the soldiers advancing. But remember that the intuition we are trying to account for here is that *only* the major's order is a cause of the soldiers advancing, not that both orders are. The appeal to quasi-dependence thus cannot help Lewis' deal with cases of trumping preemption.

How does Lewis deal with such cases, then, if appealing to the notion of quasi-dependence cannot help? Rather than further modifying his original definition of causation, Lewis (2000) offers an alternative—though closely related—definition which promises to handle all of the problematic cases of redundant causation discussed above at one fell swoop. It is to this alternative analysis that I now turn.

2.3 The 2000 analysis

Consider again the case of late preemption introduced above, a case in which, intuitively, Suzy's throw is a cause of the bottle shattering while Billy's is not. It is possible for Suzy's throw to have happened in ways that are slightly different from the way it actually happened. She could have thrown her rock at a slightly different time, at a slightly different angle, with slightly more force, etc. Following Lewis (2000, 188), call these slightly different

versions of Suzy's throw, including its actual version, 'alterations' of this event. It seems that, had any non-actual alteration of Suzy's throw occurred, some non-actual alteration of the bottle shattering would have occurred too. Had Suzy thrown her rock slightly earlier, for instance, the bottle would have shattered slightly earlier too.

Lewis proposes to use the notion of alteration introduced above to provide us with an alternative definition of causation. The first step on the way to such a definition involves the notion of influence. Consider two actual events c and e. According to Lewis (2000, 190), c *influences* e iff there is "a substantial range" $c_1, c_2, \ldots, c_n$ of "different not-too-distant alterations" of c (including c itself) and a range $e_1, e_2, \ldots, e_n$ of alterations of e "at least some of which differ [from one another]", such that $c_i \mathrel{\Box\!\!\rightarrow} e_i$ is true for any $i = 1, \ldots, n$. If c influences e in this sense then, intuitively, the manner (time, place, etc.) in which e occurs depends on the manner (time, place, etc.) in which c occurs. But influence is not yet causation. This is because, as Lewis (2000, §VIII) shows, influence is not transitive. The easy fix is, of course, to define causation as the ancestral of influence. In other words, according to Lewis' alternative definition of causation, an actual event c is a cause of another actual event e iff there is a chain of influence leading from the former to the latter.

How does this alternative definition handle cases of redundant causation? Start with late preemption. What would happen were Billy's throw altered (i.e. were a non-actual alteration of Billy's throw to occur) while Suzy's throw remains unchanged? The shattering of the bottle presumably would not change much. Whatever minute effect Billy's throw has on the shattering of the bottle—via, e.g., the negligible gravitational force exerted on the bottle by Billy's rock—would not, it seems, change much were a non-actual alteration of Billy's throw to occur. But what if one altered Suzy's throw in a similar way while leaving Billy's unchanged instead? In this case, it seems, the shattering of the bottle would change to a larger extent. This is, at any rate, what Lewis claims. As he puts it, "Influence admits of degree" (2000, 190–191) and, if we intuitively judge that Suzy's throw is a cause of the

bottle shattering while Billy's is not in cases of late preemption, it is because Suzy's throw influences the shattering to a much larger degree than does Billy's.

Lewis' treatment of cases of trumping preemption is very similar. Why do we intuitively think that what caused the soldiers to advance is the major's order and not the sergeant's? Because, whereas altering the major's order—changing it e.g. from 'Advance!' to 'Take cover!'—while holding the sergeant's fixed would have made a large difference to the soldiers' action, the reverse is not the case (since the major outranks the sergeant). The major's actual order thus has a much greater influence on the soldiers' action than does the sergeant's and this is what, Lewis tells us, explains our intuitive judgment that while the major's order is a cause of the soldiers advancing, the sergeant's is not.

Though Lewis' alternative analysis, equating causation with the ancestral of influence, seems to successfully handle the cases of redundant causation presented above, it nonetheless faces a number of difficulties. In fact, it faces too many objections and counterexamples for a survey of these to be appropriate here. I will focus on just one purported counterexample, due to Tomasz Bigaj, since it strikes me as decisive.[12] Here is how Bigaj (2012, 8) describes the case he presents to be a counterexample to Lewis' definition of causation in terms of influence:

> Let us consider one of the simplest imaginable set-ups: a railroad track that splits into two tracks, and a switch regulating the direction of train traffic. Suppose further that one of the two tracks leads to a dead end, and that at 6:00 PM a passenger train is supposed to pass this fork on the way to its destination. However, at 5:00 PM a bad guy creeps in and changes the position of the switch by moving the mechanical lever that operates the switch, so that the coming train is now directed onto the dead end track. As bad luck would have it, no one notices the change, and the train rolls full speed onto the wrong track, crashing at the end of it.

Bigaj claims, and I agree with him, that in such a case we intuitively judge that the throwing of the switch is a cause of the crash. But the throwing of the switch does not influence, in

[12] See, e.g., (Kvart, 2001), (Schaffer, 2001), (Strevens, 2003) or (Stone, 2009) for other purported counterexamples to Lewis' alternative definition.

Lewis' sense of the term, the crash of the train. There is no alteration of the way the switch was thrown that would result in an alteration in the way the train crashed. As Bigaj (2012, 9) puts it,

> You are free to imagine all sorts of alterations of the actual way the bad guy moved the lever that fall short of making this move unsuccessful—moving the lever at 5:05, moving it at 4:55, moving it with the left hand, kicking it, moving it slowly and deliberately, moving it lightning fast, etc.—and not a single one of these alterations would make the slightest difference in the way the train crashed.

Note that since an alteration of the event of the bad guy throwing the switch is, by definition, a slightly different version of this event, the bad guy *failing* to throw the switch, is not alteration of the event that actually occurred. In other words, failing to throw a switch altogether is not a way to throw a switch. If it was, then the bad guy throwing the switch would influence the crash and so, according to Lewis' alternative definition, it would be a cause of it.[13]

What lesson should one draw here? Lewis' alternative definition, just like his original definition, is inadequate. Should the inadequacy of both of Lewis' definitions lead one to abandon the hope of ever arriving at a definition of causation in terms of counterfactual dependence that is adequate? No, not necessarily. In fact, many of the accounts of causation currently on offer still endorse, as one of their central tenets, the claim that causation can be reduced to counterfactual dependence (though they do not necessarily endorse the further claim that counterfactual dependence can itself be understood in entirely non-causal terms). This is the case, for instance, of several of the accounts that employ the framework of structural equations to formulate definitions of causation. In the next section I will introduce the definition developed in (Hitchcock, 2007).

[13] See (Bigaj, 2012, §4) for an objection to the suggestion that Lewis might simply adopt the view that c is a cause of e iff there is a chain of either causal dependence or influence leading from the former to the latter.

2.4 Structural equations and counterfactuals

Consider again Bigaj's counterexample to Lewis' alternative definition of causation. It is intuitively true that had the bad guy not thrown the switch before 6:00 PM, the train would not have crashed. Since the bad guy actually threw the switch before 6:00 PM and since the train actually crashed, it is also true that the train would crash were the bad guy to throw the switch before 6:00 PM, at least if one adopts Lewis' analysis of counterfactuals (see Section 2.2).

Let me define two binary variables: M takes value 1 when the event of the bad guy throwing the switch before 6:00 PM occurs and 0 when it does not; C takes value 1 when the event of the train crashing occurs and 0 when it does not.[14] Given these two variables, one can represent the relation between the bad guy throwing the switch and the train crashing using the following equation:

$$C = M \tag{E}$$

The '=' sign in equation (E) is to be interpreted as representing not just an equality between the values of C and M but also a relation of counterfactual dependence. In other words, (E) says more than just that values of C covary with values of M. Since relations of counterfactual dependence are asymmetric, however, so must be the '=' sign in (E). Following conventions, let me stipulate that, in equations such as (E), values of the left-hand side variable (here, C) counterfactually depend on values of the right-hand side variable (here, M). Under this interpretation, (E) is nothing but a compact way of expressing the two following counterfactuals: $M = 1 \mathrel{\square\!\!\rightarrow} C = 1$ and $M = 0 \mathrel{\square\!\!\rightarrow} C = 0$. And these two counterfactuals are, of course, just the counterfactuals I assumed to be true in the previous paragraph.

Why are equations such as (E) widely referred to as 'structural equations'? Many philosophers of causation follow Lewis in identifying causal dependence with counterfactual dependence (under a non-back-tracking interpretation of counterfactuals, remember). For

[14] I will here restrict my attention to binary variables representing whether or not some event occurs.

these philosophers, equations such as (E) thus express relations of causal dependence, or what Hitchcock (2007, 504) calls facts about "token causal structure". But why call these equations 'structural' rather than 'causal' then? Because the use of equations to represent causal relationships was pioneered by statisticians and econometricians in the first half of the twentieth century, at a time when causal notions were in disrepute and when, as a result, scientists were more comfortable talking about structural relations than causal relations (see, e.g., Frisch and Waugh 1933, 390 for an early example). Even though times have changed and the majority view no longer is that causal notions are illegitimate (see, e.g., Hoover 2004 for a brief history), the adjective 'structural' has stuck. And this is why equations expressing relations of counterfactual (and causal) dependence such as (E) are widely referred to as 'structural equations'.

I should note that the use of equations to represent causal relations is not limited to philosophers who believe that causal dependence is identical with counterfactual dependence. One can use equation (E) to express the proposition that values of C causally depend on values of M without equating causal dependence with counterfactual dependence. In this case, of course, the '$=$' sign in (E) must be given an interpretation different from the one stipulated above. Michael Baumgartner (2013) and Nancy Cartwright (2014), for instance, both use structural equations to formulate definitions of causation without subscribing to the view that causal dependence is identical with counterfactual dependence.

How can one use structural equations such as (E) in order to define causation in terms of counterfactual dependence? There are many attempts to answer this question on offer. As I indicated above, I will here focus on just one such proposal, developed in (Hitchcock, 2007). And I will illustrate it using a variant of Bigaj's original case, one in which in which there is a backup bad guy who would have thrown the switch before 6:00 PM had the original bad guy failed to. As you will have guessed, this variant of Bigaj's original case is a case of early preemption.

Start by defining a *causal model* as an ordered pair $\langle \mathbf{V}, \mathbf{E} \rangle$, where $\mathbf{V}$ is a set of variables and $\mathbf{E}$ a set of structural equations relating these variables (Hitchcock, 2007, 499). In Bigaj's modified case, let $\mathbf{V} = \{U, M, B, C\}$ and $\mathbf{E} = \{(1), (2), (3)\}$. Variables M and C are defined in the same way as above: M takes value 1 when the event of the bad guy throwing the switch before 6:00 PM occurs and 0 when it does not; C takes value 1 when the event of the train crashing occurs and 0 when it does not. B is a binary variable taking value 1 when the event of the backup bad guy throwing the switch before 6:00 PM occurs and 0 when it does not. And U is binary a variable representing unmodeled causes of the event of the original bad guy throwing the switch before 6:00 PM. It takes value 1 when these causes are present and 0 when they are not. Equations (1), (2) and (3) below represent relations of causal dependence holding between these variables:

$$M = U \tag{1}$$

$$B = 1 - M \tag{2}$$

$$C = \max\{B, M\} \tag{3}$$

Equation (1) expresses the proposition that whether the original bad guy in Bigaj's story throws the switch before 6:00 PM depends on unmodeled causes.[15] Equation (2) expresses the proposition that whether the backup bad guy throws the switch before 6:00 PM depends on whether the original bad guy does. And equation (3) expresses the proposition that whether the train crashes depends on whether one of the original bad guy or the backup bad guy throws the switch before 6:00 PM.[16]

The causal model just specified is nothing but a formal representation (and not the

[15] When I use 'depends' in what follows I will mean both 'counterfactually depends' and 'causally depends'.

[16] Note that the max function in (3) is defined for all values of B and M even though it is impossible for the original bad guy and the backup bad guy to both throw the switch, i.e. move it so as to direct the train onto the track leading to a dead end (excluding a case in which they do so together). As noted in (Glymour et al., 2010, 172), these sorts of simplifications are common in accounts of causation that employ structural equations.

only possible one) of the variant of Bigaj's case I described above in plain English. It is possible to translate this formal representation into a graphical one, as illustrated in Figure 2.1. The rule for translating a causal model into what is commonly called a causal graph is

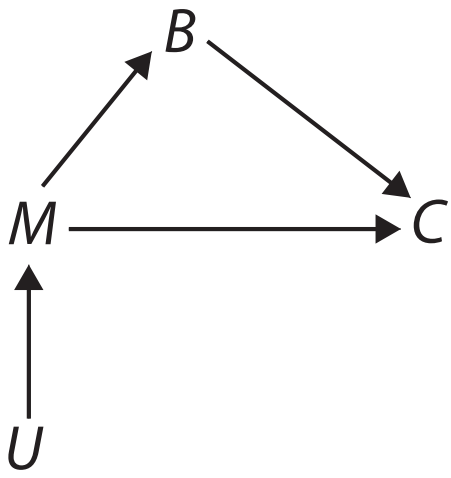

Figure 2.1: A graphical representation of the modified Bigaj case.

as follows: Draw an arrow from X into Y in a causal graph iff there is a structural equation of the form $Y = f(X\ldots)$ in the corresponding causal model. The arrows in Figure 2.1, for instance, thus represent relations of causal dependence and so, for Hitchcock, Lewis and many others, relations of counterfactual dependence. Note that graphs are less informative than the causal models they depict. Though the graph in Figure 2.1 tells you that values of B depend on values of M, for instance, it does not, unlike equation (2), give you the precise functional form of this dependence. Let me introduce a bit of terminology about causal graphs that will be useful below: If there is an single arrow from X into Y in a causal graph, then X is called a *parent* of Y and Y a *child* of X. And a series of variables $X_1,\ldots,X_n$ in a causal graph such that each X_i (for $i = 1,\ldots,X_{n-1}$) is a parent of X_{i+1} is called a *directed path* from X_1 to X_n.

The next step on the way to Hitchcock's definition of causation is to settle what it means for a counterfactual to be *true in a causal model*. Let $\langle \mathbf{V}, \mathbf{E}\rangle$ be a causal model with $\mathbf{V} = \{X_1, X_2, \ldots, X_n\}$. And let $X_i = x_i$ denote the event of X_i taking value x_i. For each X_i in $\mathbf{V}$,

E will contain an equation (E_{X_i}) of the form $X_i = f_i(X_j, \dots)$ relating it to the other variables in **V**. According to Hitchcock (2007, 501), the counterfactual $X_i = x_i \,\Box\!\!\rightarrow X_j = x_j$ is true in $\langle \mathbf{V}, \mathbf{E} \rangle$ iff X_j takes value x_j in the modified causal model $\langle \mathbf{V}, \mathbf{E}' \rangle$ one obtains by replacing (E_{X_i}) by the equation $X_i = x_i$.[17]

Though the definition just given—and which Hitchcock adapts from (Pearl, 2000, §7.1)—might appear convoluted, the underlying idea should be familiar from Section 2.2: To determine what would happen were some non-actual event to occur (respectively, were some actual event not to occur), consider a possible situation in which the event in question occurs (respectively, does not occur) while everything remains, as much as is possible, unchanged. The modified causal model $\langle \mathbf{V}, \mathbf{E}' \rangle$ is analogous to a merely possible world in which the event $X_i = x_i$ occurs thanks to a miracle, here in the form of a violation of equation (E_{X_i}).[18]

Let me illustrate Hitchcock's definition using Bigaj's modified case. What is the truth-value, in the causal model defined above, of the proposition that if the original bad guy failed to throw the switch before 6:00 PM, the train would not crash? In other words, is $M = 0 \,\Box\!\!\rightarrow C = 0$ in $\langle \mathbf{V}, \mathbf{E} \rangle$? The first step in answering this question is to replace the equation in which M appears on the left-hand side, namely (1), by the equation $M = 0$. The second step is to check the value taken by B in this modified causal model. And, as equation (2) tells you, B takes value 1 whenever M takes value 0. The third and final step is to check the value of C. Since, equation (3) tells you, the value of C is 1 whenever the value of either M or B is 1, the counterfactual considered above is false in $\langle \mathbf{V}, \mathbf{E} \rangle$. In other words, it is false that if the original bad guy failed to throw the switch before 6:00 PM, the train would not crash. This is because the backup bad guy would, in such a case, step in and thrown the

[17] Note that $X_i = x_i$ is not, strictly speaking, a structural equation and that, therefore, $\langle \mathbf{V}, \mathbf{E}' \rangle$ is not, strictly speaking, a causal model as defined above. $X_i = x_i$ does not express a relation of counterfactual dependence but, rather, expresses the stipulation that X_i is to take value x_i. You can call this equation a structural equation 'by courtesy'.

[18] See (Briggs, 2012) for a comparison of the two approaches.

switch before 6:00 PM.

Defining what it means for a counterfactual to be true relative to a causal model enables one to evaluate counterfactuals with complex antecedents—subject to the limitations discussed in (Briggs, 2012)—in a simple and mechanical fashion.[19] Does one still need Lewis' account of the truth-conditions of counterfactuals presented in Section 2.2 then? Yes, in order to account for the counterfactuals expressed by structural equations themselves. Consider again Bigaj's original case, in which there is no backup bad guy, and let the causal model for this case be $\langle \mathbf{V}, \mathbf{E} \rangle$, with $\mathbf{V} = \{C, M\}$ and $\mathbf{E} = \{(E)\}$. Should one ask whether the proposition that had the bad guy not thrown the switch before 6:00 PM, the train would not have crashed is true in this causal model? Not if what one is after is an informative answer. This is because, by calling (E) a *structural* equation and $\langle \mathbf{V}, \mathbf{E} \rangle$ a *causal* model, we are already assuming the truth of this counterfactual.[20] We still need Lewis' analysis of counterfactuals, or some suitable alternative, then, to provide truth-conditions for the counterfactuals expressed by individual structural equations.[21]

Are we now closer to Hitchcock's definition of causation? Yes, but we are not there just yet. The next step is to define what it means for *whether* one event occurs to counterfactually depend on *whether* another event occurs. This step is straightforward. Let X and Y be binary variables. And let x be the actual value of X and y that of Y. Y counterfactually depends on X in a causal model $\langle \mathbf{V}, \mathbf{E} \rangle$ iff there are non-actual values x' and y' of X and Y, respectively, such that $X = x' \boxdot\!\!\rightarrow Y = y'$ is true in $\langle \mathbf{V}, \mathbf{E} \rangle$.

[19] Of course, this procedure will only yield correct answers to queries about the truth-values of counterfactual propositions if the causal model one adopts is accurate, i.e. if it correctly depicts relations of counterfactual dependence holding 'in the world', so to speak.

[20] This would not be the case, of course, had I not assumed at the outset that the '=' in structural equations is to be understood as expressing a relation of counterfactual dependence.

[21] This does not mean that there are two distinct kinds of counterfactuals, those that are true (or false) at a world and those that are true (or false) in a causal model. For any proposition $A \boxdot\!\!\rightarrow B$, one can ask whether it is true at some world w or in some causal model $\langle \mathbf{V}, \mathbf{E} \rangle$. But if a causal model contains structural equations expressing counterfactual propositions $P_1, \ldots, P_n$, then it makes little sense to ask whether any of these propositions are true in *this* causal model, though this question might be legitimate—and its answer informative—with respect to another causal model involving different structural equations.

The next step on the way to Hitchcock's definition of causation is to characterize what it means for a variable to take a *default* value and a *deviant* value. I say 'characterize' rather than 'define' here because Hitchcock does not offer strict definitions of these two concepts. As he puts it, "the default value of a variable is the one that we would expect in the absence of any information about intervening causes." (Hitchcock, 2007, 506) But what "intervening causes" are and what we would expect to happen in their absence are dicey matters.

Consider Bigaj's modified case. It seems that the default value of C should be 0 (briefly, $Def(C) = 0$) and its deviant value 1 (briefly, $Dev(C) = 1$): In the absence of any information regarding either the original bad guy or his backup, we would expect the train to continue onto the main track and not to crash.[22] It is important to note that the default value of a variable need not be the value it actually takes. In Bigaj's modified case, for instance, the default value of C is 0 and this even though its actual value is 1. What should the default and deviant values of M and B be? This question does not seem to admit of a straightforward answer: Is the default behavior of the original bad guy to throw the switch before 6:00 PM or not to do so? This issue is one of importance since, as we will see below, the notions of default and deviant values of variables in causal models plays a central role in Hitchcock's definition of causation.

What is the next step on the way to Hitchcock's definition of causation? It is to define what it means for a set of variables in a causal model to be a *causal network* (Hitchcock, 2007, 509). Let $\langle \mathbf{V}, \mathbf{E} \rangle$ be a causal model and let $X, Y \in \mathbf{V}$. The causal network connecting X to Y in this causal model is the set $\mathbf{N} \subseteq \mathbf{V}$ that contains X, Y and all the variables on a directed path from X to Y in the causal graph corresponding to $\langle \mathbf{V}, \mathbf{E} \rangle$. In Bigaj's modified case, for instance, the causal network connecting M to C is the set $\{M, B, C\}$.

Now that the notion of causal network has been defined, one can define what it

[22] This is in agreement with the assumption made in (Hitchcock, 2007, 506).

means for such a causal network to be *self-contained* (Hitchcock, 2007, 509). Let **N** be the causal network connecting X to Y in some causal model $\langle \mathbf{V}, \mathbf{E} \rangle$. This causal network is self-contained iff, for any variable $Z \in \mathbf{N}$, if Z has parents in **N** then Z takes a default value whenever (i) its parents in **N** take their default values and (ii) its parents not in **N** take their actual values.

Instead of illustrating this definition using Bigaj's modified case right away, let me first introduce, at long last, Hitchcock's definition of causation:

> *TC*: Let $\langle \mathbf{V}, \mathbf{E} \rangle$ be a causal model, let $X, Y \in \mathbf{V}$, and let $X = x$ and $Y = y$. If the causal network connecting X to Y in $\langle \mathbf{V}, \mathbf{E} \rangle$ is self-contained, then $X = x$ is a [...] cause of $Y = y$ in $\langle \mathbf{V}, \mathbf{E} \rangle$ if and only if Y counterfactually depends on X in $\langle \mathbf{V}, \mathbf{E} \rangle$. (Hitchcock, 2007, 511)

TC differs from Lewis' definitions of causation in at least two striking ways. First, it defines causation as a relation holding between values of variables rather than between events. Second, it renders causation relative to causal models: $X = x$ can be a cause of $Y = y$ in one causal model while failing to be in another. How can one translate *TC* into a language that is closer to that used by Lewis? Here is Hitchcock's own proposal:

> [Event] c is a [...] cause of [event] e just in case: (i) $X = x$ is a [...] cause of $Y = y$ in causal model $\langle \mathbf{V}, \mathbf{E} \rangle$, as defined [by *TC*]; (ii) $X = x$ represents [the occurrence of] c and $Y = y$ represents [the occurrence of] e; and (iii) $\langle \mathbf{V}, \mathbf{E} \rangle$ is an *appropriate* causal model of the situation in which c and e occur. (Hitchcock, 2007, 503, emphasis original)

When is a causal model appropriate? At the very least, when it does not entail any false counterfactuals. Hitchcock, however, does not say much more on the topic and admits that whether a causal model is appropriate will often depend on pragmatic factors.

Let me now illustrate Hitchcock's definition of causation using Bigaj's modified case. Remember first that, in the causal model $\langle \mathbf{V}, \mathbf{E} \rangle$ corresponding to Bigaj's modified case, $\mathbf{V} = \{U, M, B, C\}$ and $\mathbf{E} = \{(1), (2), (3)\}$. Remember also that the actual values of the

variables in **V** are as follows: $U = 1$, $M = 1$, $B = 0$ and $C = 1$. Does Hitchcock's definition of causation yield the intuitively correct result that the event of the original bad guy throwing the switch is a cause of the event of the train crashing if one assumes that the causal model $\langle \mathbf{V}, \mathbf{E} \rangle$ defined above is appropriate?

The first step in answering this question is to assign default and deviant values to the variables in **V**. Above I said that the default value of C should be 0 and its deviant value 1. Let me follow Hitchcock's recommendation and stipulate that, for all variables $Z \in \mathbf{V}$, $Def(Z) = 0$ and $Dev(Z) = 1$. The second step is to identify the causal network connecting M to C in $\langle \mathbf{V}, \mathbf{E} \rangle$. It is simply the set **N** = $\{M, B, C\}$.

The third step is to determine whether this causal network self-contained. Let me start with M. Since it does not have parents in **N**, we can here ignore it. Next up is B. It has one parent in **N**, namely M. As structural equation (2) tells you, however, B takes a deviant value, namely 1, whenever M takes its default value, namely 0. In other words, B has a parent in **N** but it is not true that B takes its default value whenever this parent takes its default value. This means that the causal network **N** connecting M to C is not self-contained. And this in turn means that *TC* does not apply to it, i.e. it cannot tell you whether or not the event of the original bad guy throwing the switch before 6:00 PM is a cause of the train crashing. This result is disconcerting. We have deployed quite a bit of technical and conceptual machinery to reach the disappointing result that *TC* cannot, of Hitchcock's own admission (see e.g. Hitchcock 2007, 521), handle a very simple case of early preemption that Lewis' 1973 definition (unlike his 2000 definition) has little trouble dealing with.

Would things be different had we assigned different default and deviant values to the variables in **V**? What if, for instance, I had stipulated that $Def(B) = 1$ and $Dev(B) = 0$? Would **N** then be self-contained? As above, we can ignore M since it has no parent in **N**. If $Def(M) = 0$ and $Def(B) = 1$ then, according to structural equation (2), B does take its default value when M, its lone parent (a fortiori, its lone parent in **N**), takes its default value.

What about C? It has two parents, M and B, and both are in $\mathbf{N}$. What value does C take when M and B take their respective default values, respectively, 0 and 1? As structural equation (3) tells you, C takes its deviant value, namely 1, when $M = 0$ and $B = 1$. In other words, it is not true that C takes its default whenever all of its parents in $\mathbf{N}$ take their default values. This means that the causal network connecting M to C fails to be self-contained even if one changes the default value of B from 0 to 1. The same result obtains if one flips the default values of M and B to $Def(M) = 1$ and $Def(B) = 0$.

Now, it should be clear that, had I stipulated that $Def(C) = 1$ instead of 0, this causal network would be self-contained. What result does *TC* generate, then, in the case in which we assume $Def(M) = 0$, $Def(B) = 1$ and $Def(C) = 1$ and in which, as a result, the causal network connecting M to C is self-contained? In order to answer this question, one must determine whether C counterfactually depends on M in $\langle \mathbf{V}, \mathbf{E} \rangle$. Remember that the actual value of both M and C is 1: The original bad guy actually throws the switch before 6:00 PM and the train actually crashes. For C to counterfactually depend on M in $\langle \mathbf{V}, \mathbf{E} \rangle$, then, the counterfactual $M = 0 \; \Box\!\!\rightarrow C = 0$ must be true in $\langle \mathbf{V}, \mathbf{E} \rangle$. And for this to be the case, C must take value 0 in the modified causal model $\langle \mathbf{V}, \mathbf{E}' \rangle$ one obtains by replacing (1) by the equation $M = 0$. But structural equation (2) tells you that $B = 1$ when $M = 0$. And structural equation (3) tells you that $C = 1$ when $M = 0$ and $B = 1$. C thus does not counterfactually depend on M in $\langle \mathbf{V}, \mathbf{E} \rangle$. According to *TC*, this means that $M = 1$ is not a cause of $C = 1$ in $\langle \mathbf{V}, \mathbf{E} \rangle$.[23] In other words, under the assignment of default and deviant values specified at the beginning of the present paragraph, *TC* yields the result that the original bad guy throwing the switch before 6:00 PM is not a cause of the train crashing, contrary to our intuitions incline us to think.[24]

Is this a problem for Hitchcock? Hitchcock might argued that the stipulation that

[23]The same result obtains if one flips the default values of M and B to $Def(M) = 1$ and $Def(B) = 0$ while leaving that of C to $Def(C) = 1$.

[24]Remember that I am here assuming that $\langle \mathbf{V}, \mathbf{E} \rangle$ is an appropriate causal model for Bigaj's modified case.

$Def(C) = 1$ is in direct conflict with his statement that "the default value of a variable [as] the one we would expect in the absence of any information about intervening causes." (2007, 506) This might be true in the actual world, but what about a world in which train crashes are so common that, without any information about intervening causes, one would expect the train involved in the modified Bigaj case to crash? My intuition is that, in such a world, one would nonetheless judge that the original bad guy was the cause of *this* train crash. After all, the original bad guy threw the switch, the backup bad guy did not, and had the switch not be thrown, the train would not have crashed. Whether train crashes are very common or very rare seems irrelevant to whether the original bad guy caused this particular train to crash.

But one might also object to my having set the default value of B to 1. It is not as clear, though, that a legitimate objection can be made along those lines. What should we expect from the backup bad guy in the absence of information about interfering causes? Well, it depends what counts as an interfering cause. Is the behavior of the original bad guy an interfering cause when one considers the behavior of the backup bad guy? Rather than pondering these questions, let me propose that, at a minimum, the default values of M and B should be coherent with one another. If the default behavior of the original bad guy is to flip the switch, the default behavior of the backup bad guy is not to flip the switch, and vice-versa. In other words, if the default value of M is 0 then the default value of B should be 1, and if the default value of M is 1, then that of B should be 0.

What conclusions should one draw here? First, Hitchcock's claim that cases of preemption are cases in which there is "No counterfactual dependence in a network that is not self-contained" (2007, 529) is too hasty. As I explained above, whether this is so depends on the particular assignment of deviant and default values one picks. Second, Hitchcock (2007, 529) claims that whenever there is no counterfactual dependence in a causal network that is self-contained, "we will feel strongly compelled to say that c is not a token cause of

e." Again, this claim is too hasty. Do you feel strongly compelled to say that, in the modified Bigaj case and assuming that we live in a world in which train crashes are common, the original bad guy throwing the switch before 6:00 PM is *not* a cause of the train crashing? I certainly do not. In fact, my intuitions incline me to think that the bad guy is a cause of the train crashing in this world just as in ours.

There is a third and broader conclusion to draw here. The notions of default and deviant values play a central role in Hitchcock's account.[25] Whether *TC* applies to the modified Bigaj case and what verdict it yields when it applies depends entirely on the assignment of default and deviant values one picks for M, B and C. And yet Hitchcock says very little about these notions. This is true in (Hitchcock, 2007) but also in (Halpern and Hitchcock, 2014), a paper presenting the most recent refinement of Hitchcock's original account. Halpern and Hitchcock consider a case of early preemption isomorphic to the modified Bigaj case. This case involves an original 'assassin' (their term), a backup assassin and a victim whose drink is actually poisoned by the original assassin, but would have been poisoned by the backup assassin had the original assassin failed to. And they assume, in effect, that $Def(M) = Def(B) = Def(C) = 0$ on the grounds that "It is morally wrong, unlawful, and highly unusual" for a bad guy to be throwing a railroad switch (2014, 450). But is it so unusual for *bad* guys to throw switches and *assassins* to poison drinks? Is there any sense in which a train crash can be called morally wrong or unlawful? And would our causal intuitions be different if all of these events were morally permissible, lawful and usual? The point here is that a key notion in the account of causation offered by Hitchcock (and Halpern) is in need of further elaboration if one is to dispel the impression that the account agrees with our intuitions (when it does) only because one has carefully selected default and deviant values.

The definition of causation developed in (Hitchcock, 2007) and which I introduced

[25] Just as it does in alternative accounts developed for instance in (Menzies, 2004) or (Hall, 2007).

above is rather more complicated than either of Lewis' definitions. It struggles, however, with simple cases of early preemption that Lewis' first definition had no trouble with. Does this mean that one should be pessimistic about the possibility of adequately defining causation in terms of counterfactual dependence via the machinery of structural equations? Not necessarily. The account developed by Hitchcock is not, after all, the only one on offer. But is there any good reason to expect that we will eventually arrive at an account of causation that agrees with all of our intuitive judgments about what causes what? This is the question I examine very briefly in the next section.

2.5 Conceptual Analyses of Causation: A Pessimistic Take

Lewis' project, you will remember from Section 2.1, is to provide a conceptual analysis of causation and, *in fine*, a metaphysical reduction of causal facts. Above, in Section 2.4, I implicitly presented Hitchcock as a successor to Lewis. But the project Hitchcock is engaged in is distinct from Lewis'. His project is, to put it briefly, to provide a formal model of the process generating our intuitive judgments about causation. This does not mean that one cannot use Hitchcock's account of causation (or other accounts that employ the machinery of causal models) to further Lewis' project and this is what Ned Hall (Ms.; 2007), for instance, attempts to do (but see Hitchcock 2009). If the relation represented by the '=' sign in structural equations is counterfactual dependence as analyzed by Lewis, then a definition of causation formulated using the machinery of causal models will be a reductive definition of the kind sought by Lewis, provided that no causal notions are smuggled into the formulation of this definition.

Despite the fact that Lewis and Hitchcock are engaged in distinct projects, the methodology they follow is very similar. In fact, the structure of both (Hitchcock, 2007) and (Halpern and Hitchcock, 2014) is similar to that of (Lewis, 1973a, 2000) and of the vast majority of papers written by philosophers about causation (by which I mean, remember,

actual causation): After offering a definition of causation, the author runs through a series of test cases, including cases of preemption, to show that the definition advanced yields results that are in agreement with our intuitive judgments about causation in these. And as Halpern and Hitchcock make clear (see e.g. Halpern and Hitchcock 2014, 416, note 2), their end goal is to provide a definition of causation which agrees with our intuitive judgments *in every case*.[26]

Is there any reason to think that the methodology followed by Lewis, Hitchcock, Halpern and many others will lead to the desired results? The history of definitions of causation hardly is a cause for optimism. Philosophers tend to be very clever. As a result, whenever a definition of any concept—including causation—is advanced, the safe bet is to assume that counterexamples to it are not far behind. What is more, a group of philosophers led by Clark Glymour (Glymour et al., 2010) has recently advanced several arguments supporting the view that the methodology followed by Lewis, Hitchcock and others is unlikely to ever yield an adequate definition of causation. Though a careful examination of all of the arguments developed by Glymour et al. against what they call the "Socratic strategy" is beyond the scope of this paper, let me briefly mention two of them.

First, as Glymour et al. (2010, 169) describe it, the methodology adopted by Lewis, Hitchcock and others is one of "induction from intuitions about an infinitesimal fraction of the possible examples and counterexamples." Halpern and Hitchcock (**?**), for instance, consider fewer than a dozen test cases. As Glymour et al. (2010, §2) argue, however, the number of possible cases is very large, even when one considers just a few causes. And new problematic cases arise as soon as one considers cases with more than a few causes, especially when these causes cannot be represented by binary variables (2010, §2). If Glymour et al. are right, then there is no good reason to take the fact that a definition of causation agrees with our intuitive judgments in a handful of cases to be evidence for the

[26] At least, every possible case *in which we have clear and firm intuitions*.

adequacy of this definition.

The second argument developed by Glymour et al. questions the reliability and representative character of the intuitive judgments Lewis, Hitchcock and others use to evaluate definitions of causation. As they put it,

> All instances of the Socratic strategy that we know rely on judgments of a small group of philosophers, even for unusual cases. The presumption that philosophers' judgments in puzzling cases are or ought to be authoritative is at once comforting and unwarranted. (Glymour et al., 2010, 186)

If the definitions offered by Lewis, Hitchcock and others are supposed to be definitions of the concept of causation possessed by the members of some population (here, humans), then why think that the intuitive judgments against which to evaluate them are those formulated by Lewis or Hitchcock, rather than those that are typical in the population in question? It is concerns of this kind that motivate the recent experimental philosophy movement (see e.g. Alexander 2012). As Glymour et al. point out, however, there is a dearth of experimental work—either in psychology or in philosophy—on the very cases (e.g. cases of preemption) that Lewis, Hitchcock and others take to be test cases for their definitions. And this claim as a fortiori true of the new problematic cases which, Glymour et al. argue, arise when one considers more than just a few causes. What this means is that even if Lewis, Hitchcock or others succeeded in formulating a definition that agrees with *their* intuitive judgments in every possible case, it is not clear that what they would have arrived at would be properly seen as a definition of the concept of causation possessed by humans in general.

2.6 Conclusion

Should one abandon Lewis' project of providing a reductive definition of causation in terms of counterfactual dependence? The arguments advanced by Glymour at al., if sound, support the view that this project is unlikely to be successful. If one takes the first

of the two arguments from (Glymour et al., 2010) introduced above seriously, then one should be pessimistic about our ever succeeding in formulating a definition of causation that agrees with our intuitions in every possible case, even if we restrict our attention to cases in which we have clear and firm intuitions. Even the most sophisticated definitions on offer struggle with canonical cases (e.g. cases of early preemption), to say nothing of the myriad possible cases these definitions have yet to be confronted with. And if adopting the Socratic strategy criticized by Glymour et al. is unlikely to lead to an adequate definition of causation, then it is a fortiori unlikely to yield an adequate reductive definition of causation in terms of counterfactual dependence. This means neither that Lewis' identification of causal dependence with counterfactual dependence is erroneous nor, a fortiori, that causal notions are not closely related to counterfactual notions. It does mean, however, that a large number of philosophers currently working on actual causation are, much like Bigaj's train, headed down a dead end track.[27]

[27]Though history has shown that hitting a wall is unlikely to stop them.

Chapter 3

Interventions, Invariance and Explanatory Relevance: Not So Fast

3.1 Introduction

Why care about causal explanation? Because much of the activity of scientists, especially in the social, cognitive, and biomedical sciences, is naturally construed as an attempt to explain phenomena by discovering, describing, and measuring their causes. The interventionist account developed by James Woodward and Christopher Hitchcock (Woodward, 2000, 2003b, 2004; Woodward and Hitchcock, 2003) is among the most prominent accounts of causal explanation on offer. It is popular and influential, as evidenced by the many uses to which it has been put in recent years.[1] And this is unsurprising given the important advantages it is claimed to have over its competitors. Its advocates for instance claim that it solves Salmon's problem of explanatory relevance, a problem many take to be fatal to the Deductive-Nomological (D-N) account of Hempel and Oppenheim (1948).

After some expository work (Section 3.2), I will argue below that, despite what its

[1] To cite just one among the most recent applications, (Milkowski, 2013) resorts to the interventionist framework to develop an account of computational explanation in the cognitive sciences.

advocates claim, the interventionist account of causal explanation does not in fact solve the problem of explanatory relevance (Section 3.3). I will then examine—and ultimately rebut—six objections one might raise against my argument (Section 3.4). And I will show that, by contrast with the interventionist account, Michael Strevens's kairetic account (Strevens, 2004, 2008b) does solve the problem of explanatory relevance, provided is it tweaked in one minor way (Section 3.5). My conclusion (Section 3.6) will be that the interventionist account of causal explanation must solve what I call 'the variable choice problem' if it is to solve the problem of explanatory relevance. Since it does not, as it currently stand, solve the former problem it also does not solve the latter. As a consequence, it is inadequate as an account of causal explanation, since it does not adequately draw the line between information that is causally explanatory and information that is not.

Before proceeding, I should note that the interventionist account of causal explanation with which I will be primarily concerned below is distinct from the interventionist account of causation upon which it is based, though those two accounts are closely related, as will become clear below. The interventionist account of causal explanation is also distinct from the interventionist account of explanatory depth that is based upon it (see e.g. Hitchcock and Woodward 2003). I will briefly talk about the latter account in 4.6, but see (Weslake, 2010), (Franklin-Hall, forthcoming) or (**?**) for more complete discussions of it.

3.2 The interventionist account of causal explanation

The intuition at the root of the interventionist account of causal explanation is that to causally explain the event of a variable Y taking value y, commonly abbreviated '$Y = y$', simply is "to provide information about the factors on which it depends and to exhibit how it depends on those factors." (Woodward, 2003b, 204)[2] As Woodward also describes it, causal

[2]Both the interventionist account of causal explanation and the interventionist account of causation it is based on are formulated in terms of variables. I will here follow interventionists in talking of causal and explanatory relations as holding between variables (rather than between the features of the world these

explanation is "a matter of exhibiting systematic patterns of counterfactual dependence" between the explanandum phenomenon and its various causes (2003b, 191). How does one go about exhibiting the patterns of counterfactual dependence interventionists take to be relevant to causal explanation? By providing answers to 'what-if-things-had-been-different' questions, or w-questions for short.

What are w-questions? This question is best answered using a schematic example. Consider two binary variables X and Y each taking either value 1 or value 0 (it does not matter here what these variables represent). Assume that, in the actual world, it is the case that $X = 1$ and $Y = 1$. Assume also that one is interested in explaining the event $Y = 1$. The following is a w-question with respect to this event:

(WQ) What would the value of Y have been had the value of X been 0 instead of 1 as the result of an intervention?

There are two possible answers to this w-question, assuming the relationship between X and Y to be deterministic:

(WA_1) Had the value of X been 0 instead of 1 as the result of an intervention, then the value of Y would have been 0 instead of 1.

(WA_2) Had the value of X been 0 instead of 1 as the result of an intervention, then the value of Y would have been 1.

Let me call 'w-answers' those answers to w-questions which, like (WA_1) but unlike (WA_2), relate changes in one variable to changes—by contrast with an *absence* of changes—in another variable. As I will explain below, interventionists require that generalizations support at least one w-answer in order to be causally explanatory.

Both w-questions and w-answers refer to changes in the values of variables that result from interventions. But what are interventions? An intervention on X with respect to

variables represent). Variables will be denoted by capital letters (X, Y, Z, etc.) while particular values taken by variables will be denoted by lower-case letters (x, y, z, etc.).

Y is a manipulation that results in a change in the value of X and has an effect on the value of Y, if at all, only via its effect on the value of X. The intuition underlying the appeal to interventions is that if a change in the value of X that is induced by a manipulation that has no independent effect on the value of Y is followed (temporally) by a change in the value of Y, then the causal 'responsibility' for the change in the value of Y can be attributed to the change in the value of X. The details of Woodward's definition of 'intervention' are not important for the purpose of this paper and so I omit them here (but see Woodward 2003b, 98).

As with Hempel and Oppenheim's D-N account, generalizations play a central role in explanations according to interventionists. Indeed, they play such a central role that, in their (2003), Woodward and Hitchcock present the interventionist account not as an account of what it is for an event $Y = y$ to be causally explained—which is how it is presented in (Woodward, 2003b)—but as an account of what it is for a generalization to be causally explanatory. Although the difference between these two presentations is merely one of emphasis, it underlines the fact that, for interventionists, generalizations occupy the center of the explanatory stage. What, then, is the role played by generalizations in causal explanations according to interventionists? As I mentioned above, and as Woodward (2003b, 236) describes it, their role is to "support" w-answers. Though interventionists deliberately choose to leave it vague what they mean by 'support' (see e.g. Woodward 2003b, 279), the conditions a generalization G must satisfy in order to support a w-answer W are such that G supports W if and only if it entails it.

What conditions, then, must a generalization satisfy in order to support or entail—I will use these terms interchangeably in what follows—a w-answer and thereby be causally explanatory or, equivalently, contribute to some causal explanation? In order to answer this question, I must first explain what Woodward and Hitchcock mean by 'generalization' (2003, 181-182). On their view, there are two kinds of generality, which I will label generality$_1$ and

generality$_2$. A proposition is general$_1$ when it describes properties of more than one object. The canonical 'All ravens are black', for instance, is general$_1$. A proposition is general$_2$, by contrast, if it describes both actual and non-actual properties of a particular object. The proposition that 'Raven *r* is black but could have been pink' (where *r* refers to a particular raven) is general$_2$, while 'All ravens are black' is not. And it is generality$_2$ which matters for causal explanation according to Woodward and Hitchcock. As they put it, "The right sort of generality is [...] generality with respect to other possible properties of the very object or system that is the focus of explanation." (2003, 182) A proposition must be general$_2$ in order to entail a w-answer and be causally explanatory. It must also be 'change-relating', i.e. it must relate changes in one feature of the system it describes to changes in another one of its features (Woodward, 2003b, §5.7). The proposition 'If ravens had a shrimp-based diet instead of the diet they actually have, then they would be pink instead of black', for instance, is a change-relating generalization$_2$. It thus has the right format to support w-answers and be causally explanatory. Being a change-relating generalization$_2$, however, is only a necessary condition for supporting w-answers.

So, when do change-relating generalizations$_2$, which I will simply call 'generalizations' in what follows, support w-answers? A generalization will support at least one w-answer if and only if it is both:

(i) true, and

(ii) invariant under at least one possible testing intervention.

The label 'invariant' is often used to refer to generalizations that satisfy conditions (i)-(ii) and I will sometimes follow this custom for the sake of brevity. A generalization might, of course, support more than one w-answer. The number of w-answers a generalization supports simply is equal to the number of possible testing interventions it is invariant under.[3]

[3]Interventionists hold the view that the degree to which a generalization contributes to causally explaining an event is a function of the w-answers it entails (see 4.6 below for more). As I noted above in Section

Let me use a schematic example to illustrate conditions (i) and (ii). Consider the following generalization relating variables X and Y introduced above, assuming again that the actual value of both is 1 and that the explanandum is the event $Y = 1$:

G: $Y = X$

What G says is that the value of Y is the same as that of X, i.e. that $Y = 1$ whenever $X = 1$ and that $Y = 0$ whenever $X = 0$. Despite being an equation, G is a change-relating generalization in interventionists' sense, since it says something about non-actual values of X and Y and since it relates changes in the value of the latter to changes in the value of the former. Indeed, Woodward and Hitchcock's unorthodox view of what counts as a generalization is explicitly designed to admit equations such as G as potentially being causally explanatory.

What would it mean for G to satisfy conditions (i) and (ii)? For an equation such as G to be true is for it to be "true [. . .] of the actual values" of the variables it relates, as Woodward and Hitchcock put it (2003, 6). And since I have here assumed that the actual value of both X and Y is 1, G is indeed true of the actual values of the variables it relates (or just 'true') and so satisfies condition (i). What about condition (ii)? Testing interventions are relative to generalizations. A testing intervention on X with respect to Y relative to G is an intervention which sets X to a non-actual value and is such that G predicts that, under this intervention, Y will take a non-actual value (Woodward, 2003b, 253). An intervention that sets $X = 0$ thus is a testing intervention on X with respect to Y relative to G, since 0 is a non-actual value of X and since G predicts that, when $X = 0$, Y also takes a non-actual value, namely 0. If one assumes that, were one to set $X = 0$ by an intervention, Y would in fact take value 0, then G is invariant—i.e. would remain true—under such a testing intervention, since G is true when both X and Y take value 0. And if G is invariant under such a testing intervention, then it supports some w-answer—which is none other than (WA_1)—and thus contributes to causally explaining event $Y = 1$.

3.1, (Weslake, 2010), (Franklin-Hall, forthcoming) and [Reference omitted for blind review] develop various objections to what interventionists call their account of 'explanatory depth'.

Let me make two important remarks regarding condition (ii). First, as Woodward and Hitchcock are keen to emphasize, the interventionist account of causal explanation is "existential, rather than universal in character…" (2003, 15), since it does not require that generalizations be invariant under every possible testing intervention in order to be causally explanatory. Interventionists adopt this 'existential' view because they think that most explanatory generalizations "break down"—i.e. fail to be invariant—under some testing interventions (2003, 16). Second, the fact that, let me assume, no testing intervention on X with respect to Y relative to G has actually been carried out does not matter here. As Woodward puts it, "what matters is not whether the intervention is [. . .] carried out, but whether G would continue to hold if it *were* to be carried out." (2003b, 250, emphasis added) All it takes for G to be causally explanatory, then, is for it to be (i) true and (ii) invariant under—literally—*at least one possible* testing intervention on X with respect to Y.

Now that I have presented the interventionist account of causal explanation in some detail, let me take a step back.[4] What is it exactly that interventionists claim? One can understand their account of causal explanation as consisting of the following two theses:

(T1) One causally explains an event e if and only if one provides w-answers regarding e.

(T2) A generalization—as interventionists understand the term—contributes to causally explaining an event e if and only if it satisfies conditions (i) and (ii), i.e. if and only if it supports at least one w-answer regarding e.

The thesis I will primarily be interested below is (T1), as it forms the core of the interventionist account. Underlying (T1) is the following assumption:

(A) The information encoded in w-answers is causal information.

[4]There are several ways the account might be extended, most notably to multivariate generalizations and to the indeterministic case, but presenting these possible extensions is not necessary for the purpose of the present paper.

(A) follows straightforwardly from adopting an interventionist approach to *causation*. Consider again our binary (0, 1) variables X and Y and assume as above that the actual value of both is 1. The truth of w-answer (WA_1), i.e. of the counterfactual 'Had the value of X been 0 instead of 1 as the result of an intervention, then the value of Y would have been 0 instead of 1', entails that X is a 'direct cause' (see Woodward 2003b, 59) of Y relative to the set of variables $\{X,Y\}$. If one also assumes that all other causes of Y (should there be any) are held fixed to their actual values while the intervention setting $X = 0$ occurs, then the truth of (WA_1) also entails that $X = 1$ is an 'actual cause' (see Woodward 2003b, 77) of $Y = 1$.[5] And, because generalization G supports (WA_1), it does not simply contribute to causally explaining event $Y = 1$ but is also 'causally correct' in that it represents a genuinely causal relation—and not just a pattern of correlation—between X and Y (Woodward, 2003b, 250).

Although (T1) forms the core of the interventionist account of causal explanation, (T2) should not be seen as an optional add-on to (T1). Any proper account of causal explanation must explain how scientists—presumably the leading producers of causal explanations—actually produce such explanations. More precisely, any proper account of causal explanation must explain how it is that what scientists actually do leads them to gather the kind of information the account in question deems causally explanatory. And invariant generalizations play the key role in this part of the interventionist story regarding causal explanation. According to this story, the mathematical models scientists in many disciplines build to represent causal relations and draw inferences about them, e.g. the linear models of the form $Y = f(X_1, X_2, \ldots, X_n) + U$ that are so common in the quantitative social sciences, are just invariant generalizations by another name (when they in fact represent causal relations, of course). In other words, interventionists have a straightforward explanation regarding how scientists produce causal explanations: The models they actually build, when they are 'causally correct', support w-answers and thereby encode information that is causally

[5] As the reader will have noted, interventionists take relationships of direct causation to hold between variables and relationships of actual causation to hold between particular *values* of these variables.

explanatory.

The argument I develop below in Section 3.3 is an argument to the effect that the interventionist account of causal explanation fails to solve Salmon's problem of explanatory relevance and so fails to properly draw the line between information that is causally explanatory and information that is not. If sound, this argument supports the view that (T1) is false. And, I will claim below, the falsity of (T1) entails that of (T2).

3.3 Why interventionism does not solve the problem of explanatory relevance

According to its advocates, one of the main qualities of the interventionist account of causal explanation is that it solves what Salmon (1971, 51) calls "the problem of [explanatory] relevance".[6] To solve this problem is to give an adequate account of the relation of *explanatory relevance* which intuitively holds between an explanandum and its explanans (see e.g. Hitchcock 1995, 304). In other words, it is to adequately draw the line—by providing a set of necessary and sufficient conditions—between information that is explanatorily relevant and information that is not.[7] Most philosophers take its inability to solve the problem of explanatory relevance to be fatal to the D-N account. Consider the following argument (Kyburg, 1965, 147):

> All samples of salt (i.e. sodium chloride) that have been hexed dissolve when placed in water.
>
> Sample of salt *S* has been hexed.
>
> ∴ Sample of salt *S* dissolves when placed in water.

[6] This problem is sometimes also referred to as the 'problem of explanatory irrelevance'.

[7] You might think that trying to solve the problem of explanatory relevance thus stated, i.e. as requiring a set of necessary and sufficient conditions, is a fool's errand. It is, however, the project Woodward and Hitchcock are engaged in. And this project is one which, if the argument I develop in Section 3.5 is correct, can be completed successfully.

This argument is sound: Its premises deductively entail its conclusion and, let me assume, these premises are true. Moreover, the first of these premises is a lawlike sentence: It is universally quantified, only involves purely qualitative predicates, etc. It therefore is a law of nature, at least by Hempel and Oppenheim's standards, since it is true in addition to being lawlike. According to the D-N account, this argument thus provides an explanation of the fact that sample of salt *S* dissolves when placed in water.[8]

This verdict, however, conflicts with our explanatory intuitions. Neither the fact that *S* has been hexed nor the law relating the hexing of samples of salt to their dissolving when placed in water seem to be relevant to explaining the fact that *S* dissolves when placed in water. This is so because *S* would have dissolved when placed in water whether or not it was hexed. As Salmon (1971, 34) puts it, "Salt dissolves, spell or no spell, so we do not need to explain the dissolving of this sample in terms of a hex." Kyburg's hexed salt argument thus is commonly taken—along with Salmon's well-known John Jones argument (1971, 34)—to show that the D-N account fails to solve the problem of explanatory relevance and so to be a decisive counter-example to it.

According to its advocates (Woodward 2003b, 198-200; 2004, 48-49; Woodward and Hitchcock 2003, 19), the interventionist account preserves our explanatory intuitions because the law involved in the hexed salt argument is not invariant—where here, remember, 'invariant' is shorthand for 'satisfies conditions (i) and (ii)'—and so does not support any w-answers and is not causally explanatory. As Woodward puts it, the explanatory irrelevance of this law "may be traced to [its] failure to answer any w-questions." (2003b, 200) How does one know that this law is not invariant? According to interventionists, this law is not a generalization—and a fortiori is not change-relating—since it says nothing about non-actual properties of samples of salt that have been hexed. It thus fails to meet one of the necessary conditions for being invariant and, as a result, does not support any w-answers.

[8]Note that I use 'event' and 'fact' interchangeably in what follows.

What Salmon's argument shows is that the D-N account does not adequately distinguish information that is explanatory from information that is not, since it deems both premises of the hexed salt argument relevant to explaining its conclusion when, intuitively, they are not. The interventionist account provides a basis for this intuition, at least as far as the law figuring in this argument is concerned. This law is (causally) explanatorily irrelevant, interventionists claim, because it does not support any w-answers regarding the event described in the conclusion of the argument. As should be obvious, however, the fact that the interventionist account does not mistakenly classify the law involved in the hexed salt argument as causally explanatory does not entail that it solves the problem of explanatory relevance *in general*, i.e. that it successfully distinguishes information that is causally explanatory from information that is not. And, indeed, I argue below that it does not.

Consider the following imaginary case: I am set to perform an experiment in which I will place a sample of some mineral in water. The technician in my laboratory has prepared samples of minerals of two kinds, hexed salt, i.e. sodium chloride that has been (putatively) hexed, and diamond. Assume that I randomly pick a sample of hexed salt, that I place it in water, and that it dissolves. Naturally, I want to explain the fact that the mineral I picked dissolved when placed in water. A colleague of mine who is familiar with the interventionist account suggests using the following change-relating generalization and checking for its invariance:

F: $D = M$

Here, M is a binary variable taking value 1 when the mineral I pick is hexed salt and value 0 when it is diamond. And D is a binary variable taking value 1 when the mineral I pick dissolves in water and value 0 when it does not. The actual value of both D and M is 1. As you will have noted, F has the same format as generalization G considered above and, given this format and the actual values of D and M, F is true of the actual values of the variables it

relates and therefore satisfies condition (i).

Would F remain true under any possible testing intervention M with respect to D? For this to be the case, it must first be possible for an intervention changing the value of M from 1 to 0 to occur. What is the relevant sense of 'possible' here? As Woodward puts it, "An intervention on X with respect to Y will be 'possible' as long as it is logically or conceptually possible for a process meeting the conditions for an intervention to occur." (2003b, 132) So, is an intervention that results in my picking a sample of diamond instead of a sample of hexed salt possible in this sense? It certainly seems so: One can easily conceive of counterfactual situation in which some small change (e.g. in my brain's activity, in the movement of my hand, or in the way the samples of minerals are arranged on the table) occurs and leads me to pick a sample of diamond instead of a sample of hexed salt without having any independent effect on whether the sample of mineral I end up picking dissolves when placed in water.

So, it seems safe to assume that it is possible for an intervention setting $M = 0$ to occur. Such an intervention, moreover, is a testing intervention on M with respect to D relative to F. And F would remain true under such an intervention. This is so because the counterfactual 'Had the value of M been 0 instead of 1 as the result of an intervention, then the value of D would have been 0 instead of 1' seems to be true. In other words, it is presumably true that, had the sample I picked been a sample of diamond instead of a sample of hexed salt (as the result of an intervention), it would not have dissolved in water. Generalization F thus is invariant under some possible testing intervention in addition to being true of the actual values of the variables it relates.

Because it satisfies conditions (i) and (ii), F supports the following w-answer:

W_F: Had I picked a sample of diamond ($M = 0$) instead of a sample of hexed salt ($M = 1$) as the result of an intervention, it wouldn't have dissolved in water ($D = 0$) instead of dissolving ($D = 1$).

Because it supports W_F, F contributes to causally explaining $D = 1$, i.e. the fact that the mineral I picked dissolved when placed in water. This conclusion, however, is problematic for the interventionist account. Let me explain why.

First, given the connection between the interventionist account of causal explanation and the interventionist definition of actual causation indicated above (in Section 3.2), the fact that F is invariant under a possible testing intervention setting $M = 0$ and therefore supports W_F entails that $M = 1$ is an actual cause of $D = 1$.[9] In other words, the invariance of F under such an intervention entails that the mineral I picked being a sample of hexed salt is an actual cause of its dissolving when placed in water. The statement 'The sample of mineral I picked dissolved when placed in water because it is hexed salt', which I take to be equivalent to a claim of actual causation, however, is presumably false. Indeed, the intuition that such a claim is false seems to be what drives our judgment that the premises of the hexed salt argument are irrelevant to explaining its conclusion: The sample that was placed in water did not dissolve because it is hexed salt, it dissolved because it is salt. That it was hexed is irrelevant.[10]

Now, although this is not mandated by Woodward's (2003b, 77) definition of actual causation, an interventionist might hold the view that claims of actual causation have a contrastive structure and so that the claim entailed by W_F is the contrastive 'The mineral I picked being a sample of hexed salt *rather than a sample of diamond* is an actual cause of its dissolving when placed in water *rather than not dissolving*'. Such an interventionist might also hold that the 'because' claim that is equivalent to this claim of actual causation is the contrastive 'The sample of mineral I picked dissolved when placed in water *rather than not dissolving* because it is hexed salt *rather than diamond*'. This move, however, is of little help

[9]Let me assume here that other causes of D are held fixed to their actual values while the intervention setting $M = 0$ occurs.

[10]The satisfaction of 'being hexed salt', of course, entails the satisfaction of 'being salt'. But, as interventionists are keen to emphasize, there is a "great difference between providing a nomologically sufficient condition for an outcome and specifying what that outcome depends on." (Woodward, 2003b, 209) And, for them, only the latter activity is genuinely explanatory.

to interventionists. This is because the contrastive 'because' claim presumably is as false as its non-contrastive counterpart: The mineral I picked dissolved in water rather than ... because it is salt rather than diamond, not because it is *hexed* salt rather than diamond. Even if one assumes that the invariance of F only entails the truth of this contrastive 'because' claim, assuming it does, there remains a problem for interventionists.

A second way of describing the problem with F being invariant is to say that the w-answer it supports, i.e. W_F, though it is a true counterfactual, misleadingly suggests that the mineral I picked being hexed salt is what 'made a difference' to its dissolving when what really made the difference is its being salt, the hexing being irrelevant. In other words, the w-answer supported by F does not correctly identify the factor that made a difference to the mineral I picked dissolving upon being placed in water. It is no surprise, then, that the claim of actual causation entailed by this w-answer, whether or not one interprets it as having a contrastive structure, should strike us as false.

And here is a third way of putting the problem: The w-answer supported by F, i.e. W_F, provides one with an inefficient strategy if one's aim is to obtain a sample of mineral that will dissolve when placed in water, since it suggests that one needs to obtain is a sample of salt that has been hexed when all one needs is, in fact, a sample of salt. To say that this strategy is inefficient does not mean that it is ineffective, of course, since hexed salt will dissolve when placed in water. To put it briefly, causal explanations seek to identify difference-makers for the occurrence of the effect that constitutes the target explanandum. And when they do so appropriately, they provide one with efficient strategies for producing the effect, i.e. strategies that tell you *exactly* what you need to do so, no more and no less.

These three ways of describing the problem with F being invariant and supporting W_F are three ways of fleshing out the intuition that F and the w-answer it supports fail to identify the factor doing the 'causal work' in the dissolution of the mineral I picked and therefore ought not be seen as causally explaining the event of this dissolution. The

conclusion I draw from the argument developed above is that, insofar as it classifies F as being invariant and therefore causally explanatory, the interventionist account of causal explanation fails to solve the problem of explanatory relevance it was designed to solve. In other words, the fact that the interventionist account classifies F as invariant supports the view that (T1) is false and that providing w-answers regarding an event *e* is not sufficient to causally explain it. And if (T1) is false, so is (T2): If providing w-answers regarding an event *e* is not sufficient to causally explain it, then supporting at least one w-answer regarding *e* cannot be sufficient for a generalization to contribute to causally explaining this event. Let me now examine six objections an interventionist might raise against the argument just developed.

3.4 Objections and responses

3.4.1 Objection 1

The first objection an interventionist might raise is that in the context of the experiment I described above, the only other mineral available being diamond, the claim that the mineral I picked dissolved when placed in water because it is hexed salt is literally true—contrary to what I claimed above. And, this interventionist might continue, if this claim and the corresponding claim of actual causation (whether contrastive or not) are literally true, then it is a quality of the interventionist account that it implies that F supports w-answer W_F and therefore is causally explanatory.

Let me assume—for the sake of the argument—that this interventionist is right and that it is indeed literally true that the sample of mineral I picked dissolved when placed in water because it is hexed salt. The issue with this rejoinder is that it seems to undermine the claim that Kyburg's hexed salt argument constitutes a counter-example to the D-N account. If the intuition driving the judgment that the premises of the hexed salt argument

are irrelevant to explaining its conclusion is the intuition that no sample of mineral that is placed in water dissolves because it is hexed salt, then to assume that the sample of mineral I picked dissolved when placed in water because it is hexed salt is to contradict this intuition. And so it is to remove the basis for the judgment of irrelevance regarding the premises of Kyburg's argument. To put it differently, if being hexed salt is in fact an actual cause of the mineral I picked dissolving when placed in water, then it is unclear why the law figuring in Kyburg's argument should strike us as irrelevant to explaining the dissolution of samples of hexed salt that are placed in water. To be sure, an interventionist would maintain that this law is explanatorily irrelevant, since it is not invariant. What is unclear is that, in doing so, this interventionist would be preserving—rather than contradicting—what our explanatory intuitions seem to be under the assumption granted at the beginning of this paragraph.

The hypothetical interventionist I am here considering might ask: "OK, sure, but why exactly does this matter?" The issue here is that, if Kyburg's hexed salt argument is not in fact a counter-example to the D-N account, then it is not clear that the D-N account suffers from the problem of explanatory irrelevance it has been held to suffer from since roughly 1965. This is leaving aside, of course, Salmon's own well-known counter-example to the D-N account involving John Jones, a sexually active male who regularly consumes birth control pills and fails to become pregnant. Though limitations of space prevent me from doing so here, the argument developed in the present section can be replicated taking Salmon's argument as the starting point. In general, any counterexample to the D-N account that follows the blueprint of Kyburg's argument can be turned into a counterexample to the interventionist account by following the model provided by the argument I developed above. If the D-N account does not in fact suffer from a problem of explanatory irrelevance, however, then the interventionist account does not improve over the D-N account, as its advocates claim it does, by classifying e.g. the law figuring in hexed salt argument as explanatorily irrelevant (since this law is not explanatorily irrelevant after all).

Faced with this counter-objection, our hypothetical interventionist might simply shrug her shoulders: Maybe her account does not, after all, improve over the D-N account in this respect, but surely it remains superior to it in some other respects. It should be kept in mind, however, that the rejoinder here discussed is predicated on the assumption granted at the outset, i.e. on the assumption that it is literally true that the mineral I picked dissolved when placed in water because it is hexed salt. And the argument for this first rejoinder is only as good as the argument for this assumption is.

3.4.2 Objection 2

Objections 2 through 4 share a common thread: They are objections to the effect that there is something illegitimate about the way I chose to represent the hypothetical dissolution experiment considered above. You might have the uneasy feeling that the argument developed on the basis of this experiment involves some kind of dodgy trick: My discussion of objections 2-4 aims to unearth the reasons for this unease and, if all goes well (for me, that is), to show that it is unwarranted.

The second objection an interventionist might raise, then, is that the result that F is invariant and thereby causally explanatory is an artifact of the way I defined the variables that figure in F and, in particular, of the way I defined what it means for M to take value 0. It is true that, had I stipulated that $M = 0$ not when the mineral I pick is diamond but when it is, say, regular (i.e. not putatively hexed) salt, then F would not be invariant—since it would not, in this case, remain true under interventions that set $M = 0$—and so would not be classified as causally explanatory. And, under this redefinition of what it means for M to take value 0, it would also be the case that the mineral I picked being hexed salt is not an actual cause of its dissolving when placed in water.

There are two ways one might respond to this objection. The first is straightforward, if somewhat boring: There is nothing in the interventionist account to bar one from defining

M in the way I did above, i.e. as taking value 0 when the mineral I pick is diamond. Even though variables are the basic building blocks of their account, interventionists say very little regarding the way these variables should be defined if they are to be used in representing causal and explanatory relations.[11] They are aware, of course, of the fact that their account renders causal and explanatory claims representation-sensitive and sensitive, in particular, to the way the variables between which causal relations are supposed to hold are defined (see e.g. Woodward 2003b, 56). Indeed, in his response to Strevens's review of (Woodward, 2003b)—in a section the title of which happens to be 'Variable Choice'—Woodward (2008, §8) claims that he "assumed", in his (2003b), "that one can formulate rationally defensible non-arbitrary considerations or guidelines about which variables to employ in representing different sorts of situations–considerations that will disqualify some possible representations although they may not always pick out a uniquely best one." (2008, 211-212) He adds, however, that he "would be the first to concede that there is much more to be said on this topic." (2008, 212)

One can only agree with Woodward's assessment here. In fact, what the argument developed above shows, if sound, is that much more *must* be said on what one might call 'the problem of variable choice' if the interventionist account is to solve the problem of explanatory relevance. And neither am I alone in pointing out the importance of this problem for interventionists: In a recent paper, Franklin-Hall (forthcoming) shows how the absence of constraints guiding variable choice undermines some of the claims made by interventionists regarding the ability of their account of explanatory depth to explain the intuition that more abstract, 'high-level' explanations are sometimes better than 'low-level' explanations. If interventionists are aware that they have a problem of variable choice to solve, then, they appear to have so far underestimated the importance of what hangs on its resolution.[12]

[11] Interventionists avoid the topic of the metaphysical status of variables altogether.

[12] The problem of variable choice is, of course, a problem for any philosophical account that is formulated in terms of variables and not just for interventionism (as Franklin-Hall herself points out). But this should offer interventionists little comfort. It remains the case that the stakes are particularly high for them, since

So much for this first response—call it the 'boring response' for future reference. The second response to objection 2 is as follows: Given the set-up of the hexed salt experiment described above—in which the only two options for me to pick from are hexed salt and diamond—it seems that one *ought* to define M as taking value 0 when I pick a sample of diamond. In the context of this experiment, it seems that it would be misleading to define M as taking value 0 when I pick a sample of non-hexed salt, for instance, since there were no samples of non-hexed salt for me to pick. What this means is that, even if one assumes that interventionists have found a solution to their variable choice problem—i.e. have identified suitable constraints on the way variables should be defined—it seems that this solution ought to yield the result that defining M as taking value 0 when I pick a sample of diamond is not only legitimate but is also recommended. Objection 2, then, does not seem to hold much promise for interventionists.

3.4.3 Objection 3

The third objection an interventionist might raise takes issue with my use of definite descriptions (e.g. 'The mineral I pick') in defining the variables that figure in F (e.g. 'M takes value 1 when the mineral I pick is hexed salt'). According to this third objection, the invariance of F under some possible testing intervention is an artifact of having defined variables D and M in this way. Had I instead defined M as taking value 1 when S—some particular hunk of mineral—is a hunk of hexed salt and 0 when this very hunk is instead a hunk of diamond, and D as taking value 1 when this hunk dissolves when placed in water and 0 when it does not, F would not be invariant under any possible testing intervention. Why not? Because, our hypothetical interventionist would surely claim, there is no possible intervention that will change S, this particular hunk of hexed salt, into a hunk of diamond. If such an intervention is impossible, however, then F is not invariant and therefore does

several of their central claims (e.g. the claim that their account solves the problem of explanatory relevance) hang on the resolution of this problem.

not contribute to causally explaining the fact that S dissolves when placed in water. The impossibility of an intervention on the value of M also entails that S being a hunk of hexed salt is not an actual cause of its dissolving when placed in water.

There are three responses one might make to this third objection. The first response is simply the 'boring response' introduced above: There is no provision in the interventionist account against using definite descriptions when defining variables. The second response takes the form of a question: What principled reason is there to exclude definite descriptions from figuring in definitions of variables? It seems that we routinely—and unproblematically—use definite descriptions when describing causal relations, so why preclude these from figuring in definitions of variables that are to be used to represent causal relations? Consider a variable C representing the outcome of a coin toss and taking value 1 when the coin comes up heads and 0 when it comes up tails. There seems to be nothing wrong with defining C in the way I just did, i.e. as representing 'the outcome of a coin toss', whatever it turns out to be. Now define a variable E representing the amount of cash I have in my wallet at t, where t is some time shortly after the coin toss. And assume that I bet \$10 that the coin would come up heads. It seems natural enough to claim that C is causally relevant to E (or, in strict interventionist terms, that C is a direct cause of E relative to the set of variables $\{C,E\}$). And, assuming that the coin unfortunately comes up tails, it also seems natural to say that $C = 0$ is an actual cause of $E = a - 10$, where 'a' stands for whatever amount of cash I had in my wallet before placing the bet. It is unclear, then, what good reasons one could have for excluding definite descriptions from figuring in definitions of variables.

The third response to objection 3 questions the assumption that plays a key role in this objection, namely the assumption that there is no possible intervention that will change S, our particular hunk of hexed salt, into a hunk of diamond. One has to remember here that, for interventionists, an intervention is 'possible' as long as it is at least conceptually or logically possible. And why think that it is conceptually or, a fortiori, logically impossible

for there to be an intervention changing S into a hunk of diamond? To be sure, I do not know what such an intervention would concretely look like—otherwise I would be raiding the corner store for table salt instead of writing this paper. I do know, though, that it would presumably involve the transformation of sodium and chlorine atoms into carbon atoms, and the organization of the resulting carbon atoms into the right structure. And so I see no reason to think that such an intervention is either conceptually or logically impossible. In fact, it is unclear that an intervention of this kind is even nomologically impossible—in the sense that carrying it out would require that some laws of nature be violated. If you remain skeptical regarding even the logical possibility of such an intervention, then consider that you can replace diamond with any substance you like in the argument developed above: The conclusion will be the same as long as the substance in question does not dissolve when placed in water. If it does turn out that an intervention turning S into a hunk of diamond (or any other non-water-soluble substance of your liking) is possible, however, then F will be invariant and therefore causally explanatory even if one redefines variables *M* and *D* in the way suggested above, i.e. so as to rid it of definite descriptions. In other words, if an intervention of this kind is possible, then objection 3 simply falls apart.

3.4.4 Objection 4

The fourth objection an interventionist might raise consists in questioning the legitimacy of my having used a predicate such as 'being hexed' in defining variable *M* (where, again, 'hexed' means 'putatively hexed'). Surely, an interventionist might argue, no serious scientist would use such a predicate in describing the experimental set-up introduced earlier. There is an immediate response to this fourth objection and it is none other than the 'boring response': Interventionists do not impose any constraints regarding which predicates may figure in definitions of variables. Here, however, the history of the philosophy of science suggests an obvious candidate for such a constraint. An interventionist might require that

definitions of the values a variable X might take be of the form '$X = x$ if and only if Pa', where a is the name of some individual and, most importantly, P is a predicate referring to a *natural property* in the sense of (Lewis, 1983). Since, whatever natural properties turn out to be, the predicate 'being hexed salt' presumably does not refer to one, this requirement would outlaw variable M as I defined it and so would yield the result that F is not, after all, invariant and so is not causally explanatory.

There are three responses one might make to this fourth objection thus fleshed out. The first again takes the form of a question: What principled, non-ad hoc reason is there to require that definitions of the values a variable might take only involve predicates referring to natural properties? What exactly is defective or illegitimate about using the predicate 'being hexed' in defining M? The samples of salt involved in the experiment were, after all, hexed. To describe them as having been hexed thus is to describe them accurately. To be sure, the interventionist account of causal explanation will not properly draw the line between information that is causally explanatory and information that is not unless using a predicate such as 'being hexed' in defining the values M might take is ruled out. But this is obviously not a principled, non-ad hoc reason for requiring that definitions of the values variables might take only involve predicates referring to natural properties. So, why impose such a requirement?

The second response one might make to this fourth objection is that, as Franklin-Hall (forthcoming, §6) notes, an appeal to the distinction between natural and unnatural properties would be "in strong tension with Woodward's explicitly non-metaphysical proclivities." And the issue is not simply that a commitment to such a metaphysical distinction would rub Woodward the wrong way. As he himself puts it, "one of the attractions of the manipulationist [i.e. interventionist] account is precisely its unmetaphysical character" (2008, 194) A commitment to a metaphysical distinction between natural and unnatural properties would thus reduce the attractiveness of the interventionist account in this respect.

The third response to objection 4 is, I think, more decisive than the first two. The requirement that definitions of variables only involve predicates referring to natural properties is simply too stringent. The best currently available clinical evidence tells us that the regular consumption of statin drugs—e.g. a 20 mg tablet per day, every day before bed, for three consecutive months—contributes to lowering the level of one's 'bad' (i.e. LDL) cholesterol. Consider a continuous variable B representing the level of some individual's 'bad' cholesterol at some time t and a binary variable T taking value 1 when this individual takes a 20 mg tablet of statin per day, every day before bed, in the three-month period leading up to t. What clinical evidence tells us is that T is causally relevant to B. It thus seems that the interventionist account of causal explanation should yield the result that some generalization of the form $B = f(T, X_1, \ldots, X_n)$ is invariant and supports some w-answers relating changes in the value of T to changes in the value of B. The problem here is obvious: Any such generalization would involve a variable, namely T, that presumably violates the 'natural properties' requirement introduced above. Why? Because, however one draws the distinction between natural and unnatural properties, the predicate 'Taking a 20 mg tablet of statin per day, every day before bed, in the three-month period leading up to (some particular time) t' presumably does not refer to a natural property.[13] And T is just one example among many. The larger issue here is that many of the causal relations scientists are actually interested in, especially in the biomedical, social and cognitive sciences, involve causes and effects the representation of which using variables presumably requires one to refer to rather unnatural properties. If this is indeed the case, then adopting the 'natural properties' requirement suggested above would do interventionists more harm than good.

There might of course be other grounds for an interventionist to dismiss the use of a predicate such as 'being hexed' in definitions of variables as illegitimate. And, indeed,

[13] Indeed, if—as Lewis for instance does—one admits of 'degrees of naturalness, then the predicate involved in the definition of T seems to refer to a very unnatural property since it refers, for instance, to a particular time t.

finding such grounds—assuming this can be done in a principled, non-ad hoc fashion—seems to be the only way to prevent generalization F from being invariant since, I have argued, the alternative solutions suggested by objections 2 (defining *M* as taking value 0 not when the mineral I pick is diamond but when it is non-hexed salt) and 3 (redefining *M* and *D* so as to rid them of definite descriptions) are unsatisfactory.

3.4.5 Objection 5

I claimed earlier (in 4.2) that, had *M* been defined to take value 0 not when the mineral I pick is diamond but when it is instead non-hexed salt, generalization F would not be invariant. This is because, remember, whether or not a sample of salt has been hexed is irrelevant to whether or not it dissolves when placed in water. Does not the interventionist account, this fifth objection asks, then explain what the problem with our original generalization F consists in? This generalization seems to imply that the hexing of the salt was causally relevant to its dissolving, but interventionists have an explanation for why this strikes us as odd. It is because we know that a generalization, call it F′, relating *D* to *M*—redefined in the way described above—would not be invariant. Is this not enough for interventionists to solve the problem of explanatory relevance?

The problem is that the non-invariance of F′ only helps explain why the result that F is invariant strikes us as odd. In other words, it helps explain an undesirable consequence of the interventionist account of causal explanation but does not eliminate it. If anything, the non-invariance of F′ throws this undesirable consequence into sharper relief, since it gives an interventionist basis for the intuition that hexing is irrelevant to whether or not salt dissolves in water. In the same way, the fact that there (presumably) exists a 'good' D-N explanation for the dissolution—one that appeals to the chemical properties of salt and does not invoke hexing—does not undermine Kyburg's claim to the effect that the hexed salt argument is a bona fide D-N explanation and so does not help D-N theorists avoid the

conclusion that their account fails to solve the problem of explanatory relevance.

To put it as clearly as possible, then, the non-invariance of F′ has no bearing on whether F is invariant and does not, a fortiori, entail that F does not causally explain the event of the mineral I picked dissolving upon being placed in water. It thus cannot help interventionists avoid the conclusion that, since invariance under testing interventions is not sufficient for (causal) explanatory relevance, their account of causal explanation fails to appropriately draw the line between information that is causally explanatory and information that is not.

3.4.6 Objection 6

The sixth and final objection appeals to Woodward and Hitchcock's account of explanatory depth, i.e. to their account of what makes some causal explanations better, or *deeper*, than others. The idea here is that though F is invariant and causally explains the event of the mineral I picked dissolving upon being placed in water, there must be other generalizations which do not invoke hexing and provide much deeper explanations of this same event. In other words, though F *does* causally explain the dissolution, it does so very poorly in comparison with other generalizations that do not invoke hexing. What might such a generalization look like? An obvious candidate is the variant of F, call it F″, obtained by redefining M as taking value 1 when the mineral I pick is salt—whether hexed or non-hexed—and value 0 when it is diamond instead. It should be obvious that F″ is invariant and so, just like its cousin F, causally explains the dissolution. Does F″ provide a deeper explanation of this event than does F? In order to answer this question, let me briefly introduce the notion of explanatory depth.

As Woodward and Hitchcock understand it, explanatory depth is primarily a property of invariant generalizations—by which I mean, remember, generalizations that satisfy conditions (i) and (ii)—and only derivatively a property of the causal explanations these

generalizations are involved in. What determines how deep an invariant generalization is? The basic idea is straightforward: As Woodward puts it, "other things being equal, relationships [i.e. generalizations] that are more invariant [...] provide better [i.e. deeper] explanations." (2003b, 243) Matters become more complex, however, when one tries to spell out what the 'more' in "more invariant" means. Woodward (2003b) distinguishes three ways a generalization G might be 'more' invariant than another generalization G$'$ with respect to an explanandum $Y = y$ while Woodward and Hitchcock (2003) distinguish seven.

Thankfully, it does not matter for the purpose of this paper whether, and how, these two versions of the account of explanatory depth may be reconciled. What matters here is that, however one understands the details of this account, the depth of an invariant generalization is a function of the w-answers it entails (note: not of the *number* of w-answers it entails). Two generalizations G and G$'$ that entail the exact same w-answers about some explanandum $Y = y$ are equally deep relative to this explanandum. And if the set of w-answers entailed by G$'$ is a proper subset of the set of w-answers entailed by G, then G is deeper than is G$'$ relative to $Y = y$ (see Woodward 2003b, 260–262).[14]

Does Woodward and Hitchcock's account of explanatory depth yield the verdict that F$''$ is deeper than F? F$''$ entails one w-answer, namely the counterfactual stating that had I picked a sample of diamond instead of a sample of salt as the result of an intervention, it would not have dissolved. F also entails one w-answer, namely the counterfactual, labeled W_F in Section 3.3, stating that had I picked a sample of diamond instead of a sample of *hexed* salt as the result of an intervention, it would not have dissolved. Since these two w-answers are not equivalent, the sets of w-answers entailed by F and F$''$ are disjoint. One thus cannot assess the relative depth of F and F$''$ on the basis of the subset of relation. Since F and F$''$ entail the same number of w-answers, moreover, retreating to the view

[14]It is important to note that this account is an account of comparative—not absolute—explanatory depth. It thus does not allow for judgments of the form 'Generalization G is deep to degree d relative to explanandum $Y = y$' but only for judgments of the form 'G is deeper than G$'$ relative to $Y = y$'. It thus cannot provide a basis for the judgment that though F is causally explanatory, it is of (absolutely) shoddy explanatory quality.

according to which what matters to depth is the *number* of w-answers entailed would not help interventionists distinguish these two generalizations. It thus seems that, from the point of view of Woodward and Hitchcock's account of depth, F and F″ are explanatorily incommensurable.

One might, at this point, object that Woodward (2003b, 262) allows for comparisons between invariant generalizations entailing disjoint sets of w-answers on the basis of the relative "importance" of the w-answers entailed by each. The problem with this suggestion is that Woodward (2003b, 262) also claims that "importance" is a "subject-matter or domain-specific" matter. How is one to determine which of the w-answers entailed by F or F″ is the most important? Advocates of interventionism will no doubt claim that it is the one entailed by F″. But on what basis? The facts that hexing is explanatorily irrelevant or that, as a consequence, F″ intuitively provides a better causal explanation than does F obviously cannot form such a basis. By contrast, it seems that a straightforward case can be made for the view that the w-answer entailed by F is more important than that entailed by F″: Since the sample of mineral I picked was in fact a sample of hexed salt, the antecedent of the w-answer entailed by F gives a more accurate or precise description of this sample than does the antecedent of the w-answer entailed by F″.

Let me dispel a worry that might arise here. It may seem obvious to the reader that F″ is a better explanatory generalization than is F. Is it not the case that F″ will causally explain dissolutions involving samples of both hexed and non-hexed salt whereas F will only do so for dissolutions of the former kind? And is this difference not enough for interventionists to vindicate the claim that F″ is deeper than F? Here, it is important to remember that F and F″ are generalizations$_2$ and not generalizations$_1$ (see Section 3.2). In other words, they are general not because they describe properties of more than one object but because they describe both actual and non-actual properties of a particular object. And, for Woodward and Hitchcock, it is generality in this second sense—and only in this second sense—that

matters for causal explanation. Indeed, for interventionists, the fact that F″, unlike F, would remain true had I picked a sample of non-hexed salt instead of a sample of hexed salt has no bearing on the relative explanatory merits of these two generalizations. What matters is invariance under testing interventions. And a change from a sample of hexed salt to a sample of non-hexed salt is a testing intervention relative to neither F nor F″, since such a change will not set M—either in its original version in F or in its redefined version in F″—to its non-actual value 0 where, remember, $M = 0$ when I pick diamond.

The conclusion I draw from the discussion developed in the present subsection therefore is that, even though F″ intuitively seems to provide a better causal explanation of the dissolution than does F, the interventionist account of explanatory depth cannot provide a basis for this intuition. This is both an objection to the interventionist account of explanatory depth in itself, one to be added to those already developed in (Weslake, 2010), (Franklin-Hall, forthcoming) or (**?**), and an objection to the strategy suggested above for defending the interventionist account of causal explanation against the problem raised in Section 3.3.

3.4.7 Interim conclusion

I argued in Section 3.3 that F is invariant and therefore causally explains the event of the mineral I picked dissolving upon being placed in water. I also argued that this conclusion is problematic for the interventionist account of causal explanation since it supports the view that both (T1) and (T2) are false.

Then, in Section 3.4, I examined six objections to the argument developed in Section 3.3. These objections fall into two categories. According to objections 1, 5 and 6, though it is true that F is invariant, interventionists have the resources to explain why this conclusion is not problematic for their account. According to objections 2, 3 and 4, by contrast, the conclusion that F is invariant is both true and problematic, but interventionists can easily

modify their account of causal explanation so as to avoid this conclusion.

I argued that all six objections fail. I am inclined to think that the strategy that consists in biting the bullet, i.e. the strategy adopted by objections 1, 5 and 6, is hopeless. As I indicated at the end of 4.4, however, interventionists might be able to avoid the conclusion that F is invariant by finding legitimate grounds on which to prohibit the use of predicates such as 'being hexed' in definitions of variables. The right conclusion to draw from the argument I have developed above in Sections 3.3 and 3.4, then, is that, *as it currently stands*, the interventionist account of causal explanation fails to solve the problem of explanatory relevance, contrary to what its advocates claim.

3.5 How Strevens's kairetic account solves the problem

One might, at this point, be tempted to concede that the interventionist account faces a serious problem but argue that every account of explanation faces a similar problem. It might be that interventionists need to solve what I called the problem of variable choice in order for their account to solve the problem of explanatory relevance, but other accounts of explanation are on much the same boat: Advocates of the D-N account need an account of laws of nature which classifies the hexed salt generalization as not being a law; counterfactual theorists in the vein of (Lewis, 1986a) need a theory of events and a semantics for counterfactuals according to which the dissolution of the mineral I picked does not counterfactually depend upon this mineral being hexed salt; and so on for other accounts of explanation. All I have shown, then, seems to be that the interventionist account is no better than these other accounts. But this is not quite to show that it is worse or, a fortiori, that it should be abandoned.

There is, however, an account of explanation offer—namely Strevens's kairetic account—which succeeds where the interventionist account fail, provided one tweaks it in one minor way. According to the kairetic account (see e.g. Strevens 2004, 172), event

e_1 explains (at least partially) event e_2 just in case the former is a difference-maker for the latter. And e_1 is a difference-maker for e_2 just in case it appears in some *explanatory kernel* for e_2. What is an explanatory kernel? In order to define this notion, one must first define the notion of a causal model.[15] According to Strevens, a causal model is simply a set of propositions **M**: $\{P_1, \ldots, P_n\}$. In order for such a model to be an explanatory kernel for e_2, it must be the case that every event (or law) described by a member of **M** (i) actually occurred and (ii) is an INUS condition, in the sense of (Mackie, 1965), for the occurrence of e_2.[16] What are INUS conditions? To say that e_1 is an INUS condition for the occurrence of e_2 is to say that the former is an *Insufficient* but *Necessary* part of an *Unnecessary* but *Sufficient* condition for the occurrence of the latter. As Strevens (2004, 162) also puts it, e_1 is an INUS condition for the occurrence of e_2 just in case it is "an essential part of some set of actual conditions jointly sufficient for" the occurrence of e_2. I should note that, for Strevens as for Mackie, if e_1 is an INUS condition for e_2 then e_1 is a cause of e_2.[17]

A direct consequence of the kairetic account very briefly introduced in the previous paragraph is that if e_1 is not an INUS condition for e_2, then it will not appear in any explanatory kernel for e_2. As a consequence, it will not be a difference-maker for e_2 and will not, a fortiori, explain its occurrence (even partially). Is the event of the mineral I picked being hexed salt an INUS condition for the event of this mineral dissolving upon being placed in water? Yes, since the mineral I picked being hexed salt is an essential part of a set of conditions that are jointly sufficient for this mineral to dissolve in water.

But it is not what you might call a 'minimal' INUS condition, since the mineral I picked being hexed salt entails its being salt and since its being salt is itself an INUS condition for the dissolution. Define a minimal INUS conditions for an event e_2 as an INUS

[15] Strevens' notion of causal model differs from the one due to Hitchcock and which presented in Chapter 2.

[16] This is only a necessary condition for a causal model to be an explanatory kernel. See e.g. (Strevens, 2004, 172) for the full account.

[17] Unlike Mackie, however, Strevens does not seek to define causation in non-causal terms and takes a notion of causation as primitive. This enables him to avoid well-known counterexamples to Mackie's account.

condition for e_2 that does not entail any other INUS conditions for the same event. If one restricts membership in explanatory kernels to minimal INUS conditions thus defined, then no explanatory kernel for the dissolution of the mineral I picked will have as a member a proposition describing the event of the mineral I picked being *hexed* salt. As a result, and thanks to this minor tweak, Strevens's kairetic account does not, unlike Woodward and Hitchcock's interventionist account, imply that the mineral I picked being hexed salt explains its dissolution.[18]

3.6 Conclusion

Let me conclude very briefly, since part of my conclusions have already been stated in 4.7. Explanation has traditionally been conceived of as a relation $R(a,b)$ between an explanans a and an explanandum b. There are two ways to ensure that an account of explanation solves the problem of explanatory relevance (for some explanandum b). The first is to characterize R in such a way that the only events that satisfy R with respect to b are those that are in fact relevant to explaining it. This is what Strevens's kairetic account, tweaked in the way indicated above, does by requiring that any explanans be a minimal INUS condition for its explanandum. Imposing this requirement is a sure-fire way of excluding explanatorily irrelevant factors such as whether a sample of salt has been hexed. There is no need for advocates of the kairetic account to find grounds on which to prohibit the use of predicates such as 'being hexed' in describing potential explanantia.

I argued above, in Section 3.3, that Woodward and Hitchcock's interventionist account of causal explanation does not solve the problem of explanatory relevance in this first way: The way it characterizes the explanatory relation—as being a matter of invariance under testing interventions—is not enough, by itself, to ensure that it adequately draws the

[18]This modification of Strevens' account is not entirely unproblematic since it implies, among other things, that conjunctions of INUS conditions cannot belong to explanatory kernels. One could, of course, tweak the account further to solve this issue.

line between information that is causally explanatory and information that is not.

I then explored—in considering objections 2, 3 and 4 in Section 3.4—ways it might solve the problem of explanatory relevance in the second way, i.e. by restricting the set of potential explanantia so as to preemptively filter out any candidate which does satisfy *R* with respect to *b* but is not, intuitively, relevant to explaining *b*. Attempts to prohibit the use of predicates such as 'being hexed' from being used in definitions of variables are, in effect, attempts to solve the problem of explanatory relevance in this second way. None of the three attempts considered in Section 3.4 are successful, however.

What I called the problem of variable choice thus remains an outstanding problem for interventionists. One must therefore conclude that, contrary to what its advocates claim, the interventionist account of causal explanation fails to solve the problem of explanatory relevance.

Chapter 4

Probing the Depths of Explanatory Depth

4.1 Introduction

It is a truism that not all causal explanations are created equal and that some are better than others. But what are the factors the quality of a causal explanation depends on? It is to answer this question that James Woodward and Christopher Hitchcock (2003, see also Woodward 2003b, Chapter 6) have developed an account of what they call 'explanatory depth'. As I will explain below, this account is based upon the interventionist account of causal explanation developed by Woodward (2003b, Chapter 5; see also Woodward and Hitchcock 2003). According to the Woodward-Hitchcock (or 'WH', for short) account of explanatory depth, causal explanations are deeper, better or more powerful—Woodward and Hitchcock take the expressions 'explanatory depth', 'explanatory quality' and 'explanatory power' as synonyms—the more invariant the generalizations they involve are. For interventionists, causally explaining a phenomenon is "a matter of exhibiting systematic patterns of counterfactual dependence" between this event and its causes (Woodward, 2003b, 191). And the more invariant a generalization is, the better it helps us exhibit such patterns.

After some expository work in Section 4.2, I will argue in Section 4.3 that depth, as the WH account defines it, is just predictive power by another name and therefore is not a properly explanatory notion. I will also argue that it is not, as a result, a notion advocates of Inference to the Best Explanation should appeal to, despite what some have suggested (see e.g. Harker 2012; Mackonis 2013). Then, in Section 4.4, I will argue that the WH account conflicts with the view that causal explanations are better when they cite causes that are 'proportional' to their effects, a view many—including Woodward himself—find plausible. The conclusion I will draw from these arguments is that, if what one is looking for is an adequate account of the factors the quality of causal explanations depends on, then one should reject the WH account.

4.2 The WH account of explanatory depth

In order to introduce the WH account of explanatory depth, I must first briefly present the interventionist account of causal explanation it is based upon. This account is primarily an account of event explanation (I will use 'event', 'phenomenon' and 'fact' interchangeably in what follows), and the explananda it targets are of the form 'The event of variable Y taking value y' or, more compactly, '$Y = y$'.[1] As I mentioned above, causal explanation is, for interventionists, a matter of exhibiting patterns of counterfactual dependence between the explanandum and its causes. One does so by answering 'what-if-things-had-been-different' questions, or w-questions, about the explanandum.

What are w-questions? Consider two binary variables X and Y each taking either value 1 or value 0 (it does not matter here what these variables represent). Assume that, in the actual world, it is the case that $X = 1$ and $Y = 1$. Assume also that one is interested in

[1] Both the interventionist account of explanatory depth and the interventionist account of causal explanation are formulated in terms of variables, and I will follow interventionists here in talking of causal explanatory relations as holding between variables (rather than between the features of the world these variables represent). Variables will be denoted by capital letters (X, Y, Z, etc.) while particular values taken by variables will be denoted by lower-case letters (x, y, z, etc.).

explaining the event $Y = 1$. The following is a w-question with respect to this event:

(WQ) What would the value of Y have been had the value of X been 0 instead of 1 as the result of an intervention?

An intervention on X with respect to Y is a manipulation that results in a change in the value of X and has an effect on the value of Y, if at all, only via its effect on the value of X. The intuition underlying the appeal to interventions is that if a change in the value of X that is induced by a manipulation that has no independent effect on the value of Y is followed (temporally) by a change in the value of Y, then the causal 'responsibility' for the change in the value of Y can be attributed to the change in the value of X. The details of Woodward's definition of 'intervention' are not important for the purpose of this paper and so I omit them here (but see Woodward 2003b, 98).

There are two possible answers to (WQ), assuming the relationship between X and Y to be deterministic:

(WA_1) Had the value of X been 0 instead of 1 as the result of an intervention, then the value of Y would have been 0 instead of 1.

(WA_2) Had the value of X been 0 instead of 1 as the result of an intervention, then the value of Y would have been 1.

Let me call 'w-answers' those answers to w-questions which, like (WA_1), relate changes in one variable to changes—by contrast with an *absence* of changes—in another variable. As I am about to explain, interventionists require that generalizations entail at least one w-answer in order to be causally explanatory.[2]

Consider the following generalization relating our binary (0, 1) variables X and Y, assuming again that the actual value of both is 1 and that the explanandum is the event

[2]Woodward uses 'support' where I use 'entail' and deliberately chooses to leave it vague what he means by 'support' (see e.g. Woodward 2003b, 279). As will become clear below, however, the conditions a generalization G must satisfy in order to support a w-answer W are such that G supports W if and only if it entails it (see below note 4). I will thus stick to the more precise 'entail' in what follows.

$Y = 1$:

$$Y = X \tag{G}$$

G—and, more generally, equations of the form $Y = f(X)$—should be read as implicitly prefixed with a universal quantifier ranging over values of X. What G says, then, is that for any value x of X, Y takes a value y that is such that $x = y$. In other words, G says that $Y = 1$ whenever $X = 1$ and that $Y = 0$ whenever $X = 0$.[3] According to interventionists (see e.g. Woodward and Hitchcock 2003, 6), G causally explains event $Y = 1$ if and only if it is both:

(i) true of the actual values of X and Y, and

(ii) invariant under at least one possible testing intervention on X.

A generalization that satisfies conditions (i) and (ii) will entail at least one w-answer, as I will illustrate below. A generalization might, of course, entail more than one w-answer, a feature the WH account of explanatory depth will exploit. Note that the label 'invariant' is often used to refer to generalizations that satisfy conditions (i)-(ii) and I will sometimes follow this custom for the sake of brevity.

What would it mean for G to satisfy conditions (i) and (ii)? Start with condition (i), the condition requiring that G be "true [...] of the actual values" of the variables it relates, as Woodward and Hitchcock put it (2003, 6). Here, since I have assumed that the actual value of both X and Y is 1, G is indeed true of the actual values of the variables it relates (or just 'true', for short) and it therefore satisfies condition (i). What about condition (ii)? Testing interventions are relative to generalizations. A testing intervention on X with respect to Y relative to G is an intervention which sets X to a non-actual value and is such that G predicts that, under this intervention, Y will take a non-actual value (Woodward, 2003b, 253). An intervention that sets $X = 0$ thus is a testing intervention on X with respect to Y

[3] Though G is an equation, it is also a 'change-relating generalization' in Woodward and Hitchcock's terms (see e.g. Hitchcock and Woodward 2003, 181) and so has the right format to potentially be causally explanatory.

relative to G, since 0 is a non-actual value of X and since G predicts that, when $X = 0$, Y also takes a non-actual value, namely 0. If one assumes that, if one were to set $X = 0$ by an intervention, then Y would in fact take value 0, then G is invariant under such a testing intervention, since G is true when both X and Y take value 0. To say that G is invariant under an intervention setting $X = 0$ here is simply to say that it would remain true were such an intervention to occur. And if G is invariant under a testing intervention setting $X = 0$, then it entails some w-answer—namely (WA_1)—and thus causally explains event $Y = 1$.[4]

Now that I have presented the essential elements of the interventionist account of causal explanation—it can easily be extended to multivariate generalizations and to the indeterministic case—it is time to turn to the WH account of explanatory depth that is based upon it. Explanatory depth is primarily a property of invariant generalizations—by which I mean, remember, generalizations that satisfy conditions (i) and (ii)—and only derivatively a property of the causal explanations these generalizations are involved in. What determines how deep an invariant generalization is? The basic idea is straightforward: As Woodward puts it, "other things being equal, relationships [i.e. generalizations] that are more invariant [...] provide better [i.e. deeper] explanations." (2003b, 243) Matters become more complex, however, when one tries to spell out what the 'more' in "more invariant" means. Woodward (2003b) distinguishes three ways a generalization G might be 'more' invariant than another generalization G$'$ with respect to an explanandum $Y = y$ while Woodward and Hitchcock (2003) distinguish seven.

Thankfully, it does not matter for the purpose of this paper whether, and how, these two versions of the WH account of explanatory depth may be reconciled. What matters here is that, however one understands the details of this account, the depth of an invariant

[4]The invariance of G under an intervention setting $X = 0$ guarantees that it entails (WA_1) simply because the truth of (WA_1) is a necessary condition for G to be invariant under such an intervention. If (WA_1) was false and it was not the case that Y would take value 0 were the value of X to be set to 0 by an intervention, then G would not remain true under an intervention setting $X = 0$. Invariant generalizations thus trivially entail the w-answers which they "support", to use Woodward's term.

generalization is a function of the w-answers it entails (note: not of the *number* of w-answers it entails). Two generalizations G and G′ that entail the exact same w-answers about some explanandum $Y = y$ are equally deep relative to this explanandum. Moreover, if the set of w-answers entailed by G′ is a proper subset of the set of w-answers entailed by G, then G is deeper than is G′ relative to $Y = y$ (see Woodward 2003b, 260–262). It is important to note that the WH account is an account of comparative—not absolute—explanatory depth. It thus does not allow for judgments of the form 'Generalization G is deep to degree *d* relative to explanandum $Y = y$' but only for judgments of the form 'G is deeper than G′ relative to $Y = y$'.

Let me assuage a worry that might arise from the many caveats Woodward and Hitchcock attach to their account of explanatory depth. As they are keen to emphasize, "explanatory depth is not one-dimensional" (2003, 188) but, rather, is a "complicated and multidimensional" notion, as Woodward (2003b, 265) puts it. One might be tempted to read these claims as implying that the depth of a generalization is a function of other factors besides the w-answers it entails. This interpretation, however, is misguided: Each one of the dimensions referred to by Woodward and Hitchcock corresponds to one of the seven ways a generalization might be 'more' invariant distinguished in (Hitchcock and Woodward, 2003). And all of these make the depth of a generalization a function of the w-answers it entails. As Brad Weslake puts it in a recent paper, these seven ways of greater invariance are all ways in which "a generalization can provide the resources to describe a greater range of true counterfactuals concerning possible changes to the system in question—that is, to answer more *w-questions*." (2010, 278, emphasis original) It should be noted, moreover, that if this was not the case—i.e. if a generalization's depth did depend on other factors than the w-answers it entails—then there would be no such thing as the WH account of explanatory depth. Woodward and Hitchcock would have given us, at best, an account of but one component of explanatory depth, which is clearly not what they take themselves to

be doing.

Now that the WH account of explanatory depth has been introduced, let me turn to two arguments to the effect that it is inadequate as an account of the factors the quality of causal explanations depends on.

4.3 Explanatory depth and inference to the best explanation

Advocates of Inference to the Best Explanation (IBE) hold the view that, when faced with a set of incompatible hypotheses $H_1, \ldots, H_n$ all of which fit the available empirical evidence E equally well (or are empirically equivalent), one should infer to the truth of the hypothesis that best explains this evidence (see e.g. Harman 1965, 90–91). A full-fledged defense of IBE requires that one specify the criteria by which to assess the relative explanatory qualities of $H_1, \ldots, H_n$. In other words, it requires an account of explanatory virtues. David Harker (2012) has recently suggested that explanatory depth, as defined by the WH account, is an explanatory virtue advocates of IBE can appeal to (see also Mackonis 2013).[5] His concern is to identify what he calls "peculiarly explanatory virtues" (Harker, 2012, §1)—i.e. virtues of hypotheses that differ from their theoretical virtues (simplicity, fruitfulness, predictive power, ontological heterogeneity, etc.)—and this in order to prevent IBE from collapsing into the view that, when faced with empirically equivalent incompatible hypotheses, one should infer to the truth of the hypothesis that possesses the 'best mix' of these theoretical virtues.

Harker's endeavor seems sensible: If IBE is to be a distinct mode of inference, then one must be able to identify peculiarly explanatory virtues. Unfortunately, explanatory depth

[5] Because depth, as characterized by the WH account, is a property of *causal* explanations, to appeal to it in a defense of IBE is to restrict the scope of IBE to causal hypotheses, i.e. to hypotheses that posit the existence of certain causal relations.

as defined by the WH account is not such a virtue. It is simply predictive power in disguise, contrary to what Harker (2012, §3) himself claims. I will substantiate this claim below, but let me here note two of its implications: First, and obviously, if this claim is true then the notion the WH account is an account of cannot be appealed to by advocates of IBE. Second, if this notion is just predictive power, then why accept Woodward and Hitchcock's claim that it is an explanatory notion? IBE in fact provides us with a guide, albeit an imperfect one, for assessing whether a virtue of hypotheses or generalizations is explanatory. Can advocates of IBE appeal to this virtue without IBE thereby collapsing into some other mode of inference, e.g. inference on the basis of theoretical virtues? If yes, then we have a prima facie reason to classify this virtue as explanatory. If not, however, then we have a prima facie reason not to classify this virtue as explanatory.

Let me provide an illustration by way of evidence for the claim that depth, as the WH account defines it, is just predictive power. Consider two continuous variables W and Z—which might, for instance, represent two physical quantities—and the two following generalizations relating them, where $f \neq g$:

$$W = f(Z) \tag{G_1}$$

$$W = g(Z) \tag{G_2}$$

Assume now that we have a data set composed of 50 observations $\{(z_1, w_1), \ldots, (z_{50}, w_{50})\}$ and that both G_1 and G_2 are 'true' of these observations, i.e. that $w_i = f(z_i) = g(z_i)$ for any $i \in [1, 50]$. Given these two assumptions, G_1 and G_2 are empirically equivalent.

At this point, neither G_1 nor G_2 causally explains any of the 50 observations according to the interventionist account of causal explanation. In order for these generalizations to causally explain one of these observations, they must both entail at least one w-answer regarding the observation in question. Let me here stipulate that the target explanandum is

$W = w_1$. If one also assumes both the truth of the following counterfactual:

(1) Had the value of Z been z_{51} instead of z_1 as the result of an intervention, then the value of W would have been w_{51} instead of w_1.

and also that $w_{51} = f(z_{51}) = g(z_{51})$, then both G_1 and G_2 are invariant under at least one possible testing intervention. As a result, they both entail one w-answer, namely (1), and so causally explain the event $W = w_1$.[6] But notice that the fact that both G_1 and G_2 entail (1) simply means—as a matter of definition—that they both accurately predict what would happen were the value of Z set to $Z = z_{51}$ by an intervention.[7] We have yet to observe a situation in which Z takes value z_{51} as the result of an intervention. And both G_1 and G_2 tell us—here, accurately—which value of W we should expect to observe in such circumstances. This is just prediction.

The step to the claim that depth, as defined by the WH account, is just predictive power is a short one. Assume the truth of the following counterfactual:

(2) Had the value of Z been z_{52} instead of z_1 as the result of an intervention, then the value of W would have been w_{52} instead of w_1.

And assume also that $w_{52} = f(z_{52}) \neq g(z_{52})$. Given these two assumptions, G_1 entails w-answer (2) while G_2 does not. The set of w-answers entailed by G_2 thus is a proper subset of the set of w-answers entailed by G_1. As a result, G_1 is explanatorily deeper than G_2 according to the WH account, i.e. the former provides us with a better causal explanation of event $W = w_1$ than does the latter. But this simply means that G_1 provides us with more true predictions regarding the value of W—and the way it changes under interventions on

[6]Note that the fact that $w_2 = f(z_2) = g(z_2)$, for instance, does not imply that either G_1 or G_2 entail the following counterfactual: (2) 'Had the value of Z been z_2 instead of z_1 as the result of an intervention, then the value of W would have been w_2 instead of w_1'. It might be that, as a matter of actual fact, $w_2 = f(z_2) = g(z_2)$ but that this equality would not hold in counterfactual circumstances in which Z comes to take value as z_2 as the result of an intervention (which need not have actually been the case).

[7]Note that I use 'predict' in a temporally neutral sense here, so that it covers both prediction as commonly understood (about the future) and retrodiction (about the past).

Z—than does G_2. G_1 is deeper than G_2 simply in virtue of the fact that it correctly predicts the value W would take under an intervention setting $Z = z_{52}$ while G_2 does not, i.e. simply in virtue of the fact that it has greater predictive power.

Here is how Harker describes what he calls the "more ambitious" version of the strategy that consists in advocates of IBE appealing to depth as defined by the WH account: "If pursuing invariance helps us achieve deeper explanations, for example, and deeper explanations indicate a more truthlike theory, then we connect a distinctively explanatory virtue to perhaps the ultimate scientific achievement." (2012, §3) The fact that depth, as the WH account defines it, is just predictive power by another name is both good and bad news for those who hope to follow this strategy. It is good news because the connection between depth thus defined and truthlikeness is seemingly straightforward. Remember that generalizations G_1 and G_2 are to be read as implicitly prefixed with a universal quantifier ranging over values of Z. The fact that G_1 entails w-answers (1) and (2) while G_2 only entails (1) means that G_1 is true of more values of Z than is G_2. While G_1 is true of 52 values of Z, G_2 is true of 51. In this sense, G_1 is closer to the truth than is G_2, i.e. it is closer to being true of all the values of Z. It should also be clear that, in this sense of 'truthlike', the deeper a generalization is, the more truthlike it is. Mechanically, greater depth guarantees greater truthlikeness, but trivially so: Greater depth simply is greater truthlikeness.

The bad news for advocates of IBE who wish to appeal to depth as defined for the WH account is obvious: As I noted above, if IBE is the view that, when faced with empirically equivalent incompatible hypotheses (or generalizations), one should infer to the truth of the deepest one, then IBE simply collapses into the view that one should infer to the truth of the hypothesis that is closest to the truth. It is presumably true, if not trivially true, that, when faced with empirically equivalent incompatible hypotheses, one ought to infer to the truth of the one that is closest to the truth (at least if one is a scientific realist). But IBE was supposed to give us an independent purchase on truthlikeness via explanatory depth or

power. It should be clear that, if one weds IBE and depth as defined by the WH account, then IBE cannot deliver on its promise.

As I indicated above, there is a broader conclusion to draw here: If depth as the WH account defines it is just predictive power, then why think that this account is adequate as an account of the factors the quality of causal *explanations* depends on? And why think that the interventionist account the WH account of depth is based on is properly seen as an account of causal *explanation*? To be sure, the predictions entailed by invariant generalizations are predictions of a specific kind, namely predictions regarding what would happen under various interventions. But predicting what would happen under various interventions is still predicting. I am not here claiming that there is no relationship at all between prediction and explanation. I am simply questioning the view adopting the WH account commits one to, namely the view that causal explanation and what one might call 'causal prediction' are one and the same thing.[8] At any rate, if you think—as most contemporary philosophers of science do—that prediction and explanation are conceptually distinct, then you should have serious doubts regarding Woodward and Hitchcock's claim to have provided accounts of causal *explanation* and *explanatory* depth.

I hope to have done enough here to convince the reader (i) that there are good reasons to be skeptical of Woodward and Hitchcock's claim to be providing accounts of explanatory, as opposed to predictive, notions and (ii) that, as a result, depth as the WH account defines it is not a virtue advocates of IBE should appeal to. Let me now turn to a distinct shortcoming of the WH account of depth, one that subsists even if one bites the

[8]The view that prediction and explanation, whether causal or not, are the same thing was famously endorsed by Carl Hempel and Peter Oppenheim, who claimed that the distinction between the two notion "is of a pragmatic character." (1948, 138) This view has fallen out of favor for a variety of reasons. But note that, regardless of what one thinks of this view, there is an important difference between Hempel and Oppenheim on the one hand and Woodward and Hitchcock on the other. According to Hempel and Oppenheim's Deductive Nomological (DN) account of explanation, explaining an event such as $W = w_1$ is the same thing as predicting (or retrodicting, assuming that we normally seek to explain events that have already occurred) the occurrence *of this very event*. By contrast, for Woodward and Hitchcock, causally explaining $W = w_1$ is the same thing as predicting the occurrence *of some other event*, e.g. of event $W = w_{52}$.

bullet and follows Woodward and Hitchcock in identifying causal explanation with causal prediction and explanatory depth with predictive power.

4.4 Explanatory depth and proportionality

Many philosophers—including Woodward himself (see e.g. Woodward 2010)—believe that causal explanations are better when they cite causes that are 'proportional', in the sense of (Yablo, 1992, 277), to their effects. Indeed, some—e.g. Christian List and Peter Menzies (2009)—even think that proportionality is necessary for causation. The case canonically used to introduce Yablo's notion of proportionality is that of Sophie, a pigeon who pecks at all and only red targets (Yablo, 1992, 257). Imagine that, on some occasion, I present Sophie with a crimson target and, crimson being a shade of red, she pecks at it. What is it that caused Sophie to peck at the target? There are at least two ways to answer this query for a causal explanation. The first is to say that the cause of Sophie's pecking was the target being red. The second is to say that it was the target being crimson.

According to advocates of proportionality, the explanation that mentions the target being red as the cause of Sophie's pecking is the better one because it cites a cause that is proportional to its effect. The explanation that cites the target being crimson cites a cause that is too 'specific' for its effect since, by assumption, Sophie would have pecked at the target had it been red but non-crimson.[9] To put it briefly, a cause is proportional to its effect in Yablo's sense just in case it is both necessary and sufficient for its effect. Being presented with a red target is both necessary and sufficient for Sophie to peck since, as I stipulated above, Sophie pecks at all and only red targets she is presented with. Being presented with a crimson target, by contrast, is sufficient but not necessary for Sophie to peck, since she will peck at red but non-crimson targets.

[9]For advocates of the view that proportionality is necessary for causation, e.g. List and Menzies, this second explanation is not even properly called 'causal' insofar as, on their view, the target being crimson is not a cause of Sophie's pecking, since it violates the proportionality requirement.

What does this all have to do with explanatory depth as defined by the WH account? As I will show below, there are cases in which the WH account implies that a generalization citing a cause that is not proportional to its effect is just as deep as a generalization that does cite a proportional cause. As a result, one cannot consistently endorse both the WH account of depth and the view that causal explanations are better when they cite causes that are proportional to their effects.

Suppose that you are seeking to causally explain the fact that some sample of water located in a room the temperature of which is below 0°C is in a frozen state (this case is adapted from Craver 2007, 205–206). And assume that the two following generalizations are available:

$$S = C \tag{G$_3$}$$

$$S = \begin{cases} 0 & \text{if } T \leq 0 \\ 1 & \text{if } T > 0 \end{cases} \tag{G$_4$}$$

Here S is a binary variable taking value 0 when the sample of water is frozen and value 1 otherwise (i.e. if it is either liquid or gaseous). Variables C and T embody alternative ways of representing the temperature of the room the sample of water is located in. C is a binary variable taking value 0 when the temperature is less than or equal to 0°C and value 1 when it is strictly greater than 0°C. And variable T is a discrete variable taking its values in $\mathbb{Q}$ and representing the temperature of the room on the Celsius scale.

I assumed above that the sample of water of interest is in a frozen state. This means that the actual value of S is 0. Let me also stipulate that the temperature in the room is -19.6°C, so that the actual values of C and T are 0 and -19.6, respectively. Both G$_3$ and G$_4$ are thus true of the actual values of the variables they relate. Assume now that the two following counterfactuals are true:

(3) Had the value of C been 1 instead of 0 as the result of an intervention, then the value of S would have been 1 instead of 0.

(4) Had the value of T been 10 instead of -19.6 as the result of an intervention, then the value of S would have been 1 instead of 0.

The truth of (3) straightforwardly implies that G_3 is invariant under at least one testing intervention. As a result, G_3 entails a w-answer, namely (3) itself, and therefore causally explains the event $S = 0$. The truth of (4) likewise implies that G_4 is invariant under at least one testing intervention. G_4 thus entails w-answer (4) and causally explains the event $S = 0$.

As should be obvious from the way C and T were defined above, the following relations hold between values of these two variables:

$$C = 0 \leftrightarrow T \leq 0 \qquad (B_1)$$

$$C = 1 \leftrightarrow T > 0 \qquad (B_2)$$

Given B_1 and B_2, any w-answer entailed by G_3 will also be entailed by G_4, and vice-versa. Imagine that somebody asks you 'What would the value of S be were T to be set to 10 by an intervention?' Since, by B_2, any intervention that sets $T = 10$ also sets $C = 1$, you can answer this w-question without knowing G_4, simply with the help of G_3. And if you rely on G_3 in order to answer this query, the answer you will give will be none other than w-answer (4), since G_3 tells you that $S = 1$ when $C = 1$. Likewise, you can answer the w-question 'What would the value of S be were C to be set to 1 by an intervention?' without knowing G_3, simply with the help of G_4. The answer you will then give will be w-answer (3), since any intervention that sets $C = 1$ also sets T to a value > 0 and since G_4 tells you that $S = 1$ when $T > 0$.

Thanks to the relationships holding between values of C and values of T, generalizations G_3 and G_4 thus entail exactly the same w-answers with respect to explanandum $S = 0$. According to the WH account, this means that they provide us with equally good causal explanations of the fact that the sample of water is frozen. The problem here is that,

for advocates of the view that causal explanations are better when they cite causes that are proportional to their effects, G_3 provides us with a better causal explanation of the fact that $S = 0$ than does G_4. This is because G_3 cites a cause that is proportional to its effect while G_4 does not. The cause cited by G_4, namely the exact temperature of the room the sample of water is located in, is intuitively too 'specific' insofar as the target explanandum is the relatively coarse-grained fact that the sample of water is frozen. The analogy with the case of Sophie, Yablo's pigeon, is straightforward. In Yablo's case, we know—because we stipulated it—that Sophie will peck if and only if the target is red. In the case at hand, we know that the sample of water will be frozen if and only if the temperature in the room is less than or equal to $0°$C. To cite the precise shade of the red target Sophie was presented with, namely crimson, as the cause of her pecking is to cite a cause that is too 'specific' because Sophie would have pecked at a red but non-crimson target. And, here, to cite the precise temperature of the room, namely $-19.6°$C, is to cite a cause that is too 'specific' because the water would have been frozen had this temperature been different from $-19.6°$C as long as it was less than or equal to $0°$C.

The WH account and the proportionality view thus yield conflicting verdicts. According to the former, G_3 and G_4 provide us with equally good causal explanations of the fact that the sample of water is frozen. And according to the latter, G_3 provides us with a better causal explanation than does G_4. One thus cannot consistently endorse both the WH account and the proportionality view. If you think, as many (but by no means all) philosophers do, that the proportionality view is correct—and that G_3 provides us with a better causal explanation because it tells us exactly what we need to know, no more and no less, to causally explain the coarse-grained fact that $S = 0$ while G_4 tells us too much—then the argument developed above, if sound, provides you with a good reason to reject the WH account of explanatory depth.

4.5 Conclusion

I have argued above that depth as the WH account defines it is not an explanatory notion but, rather, is simply predictive power by another name. As I explained, this means that this notion is not one advocates of IBE looking for what Harker calls "peculiarly explanatory virtues" should appeal to. I have also argued that the WH account conflicts with the popular view that causal explanations are better when they cite causes that are proportional, in Yablo's sense, to their effects. If I am right that depth as the WH account defines it is just predictive power, then this is hardly surprising. If proportionality is a norm governing causal explanation while depth as defined by the WH account is a norm governing what I have called 'causal prediction' then, given that prediction and explanation are conceptually distinct activities, there is no reason to expect these two norms to agree. If what you are looking for is an adequate account of the factors the quality of causal explanations depends on, then, there are good reasons for you to reject the WH account of explanatory depth.

Chapter 5

Interventionism Does Not Explain the Practical Usefulness of Causal Knowledge

5.1 Introduction

James Woodward (2014) advocates a "functionalist" approach to causation. If you want to evaluate a definition of causation, ask not whether it is reductive or how closely it fits our intuitive causal judgments. Instead, ask whether it can help you make sense of the various functions causal knowledge plays in our lives. One of these functions—perhaps the main one—is that of guiding our decision-making by helping us predict the effects of our actions. As Woodward himself puts it, "Causal knowledge is knowledge that is useful for a very specific kind of prediction problem: the problem an actor faces when she must predict what would happen if she or some other agent were to act in a certain way…" (2003b, 32)

The view that causal knowledge—by contrast with knowledge of mere correlations—is practically useful because it helps us predict the effects of our actions is not, of course,

original to Woodward. It has been the received view in the contemporary philosophical literature at least since Nancy Cartwright's seminal "Causal Laws and Effective Strategies" (1979). But Woodward also claims that his popular and influential interventionist account of causation, developed at length in (Woodward, 2003b), provides an explanation for the practical usefulness of causal knowledge thus understood, i.e. a bridge leading from causal claims to predictions regarding the effects of our actions. And, indeed, providing such an explanation is often taken to be one of the main qualities of the interventionist account (see e.g. Kuorikoski 2014, 337-338). I will argue below that, despite appearances—and claims—to the contrary, Woodward's account fails to explain why causal knowledge should be useful for predicting the effects of our actions, and this even if one assumes that it is correct as an account of causation. If we are to judge the interventionist account by the functionalist standard Woodward advocates, we should conclude that it is lacking in an important way.[1]

According to Woodward's interventionist account, causal claims of the form 'X is a cause of Y' (where X and Y are variables) are equivalent to counterfactual claims regarding what would happen to the value of Y were an *intervention* on X with respect to Y to occur. An intervention on X with respect to Y is, to put it briefly, a manipulation of the value X which has an effect on Y, if at all, only via its effect on X (see Woodward 2003b, 98 for the details). The connection between interventions and causation is motivated by the following thought: If a change in the value of X that is induced by an intervention on X with respect to Y is temporally followed by a change in the value of Y, then the change in the value of X must be responsible for the change in the value of Y.

Causal claims thus are, for interventionists, equivalent to counterfactual claims regarding what would happen were certain kinds of manipulations to be carried out. The

[1]Note that I am here concerned only with the views defended by Woodward, at the exclusion of those advocated by Pearl (2000) or Spirtes, Glymour and Scheines (2001). These views, especially Pearl's and Woodward's, are sometimes lumped together—inappropriately in my view—under the label 'interventionist' (see e.g. Kuorikoski 2014, 334).

relevance of causal claims thus analyzed to prediction problems faced by decision-makers seems obvious. If the causal claim 'Taking aspirin relieves fevers' is, as a matter of definition, equivalent to a counterfactual regarding what would happen to one's fever were one to take aspirin, then it is no wonder—it seems—that knowledge of this causal claim should be useful for predicting what would happen were one to take aspirin and should therefore guide one's decision whether to take aspirin, given that one has a fever.

There is, however, a serious issue with this apparently simple and elegant explanation for the practical usefulness of causal knowledge: It does not apply to a great many cases in which causal knowledge is in fact practically useful and helps us predict the effects of our actions, including the fever-and-aspirin case. To see why, let's look at this case in more detail.

5.2 The set-up: Aspirin and fevers

The causal structure represented in Figure 5.1 is, I take it, quite typical of situations in which one has a fever and is pondering whether to take aspirin. Here H_{t_1} and H_{t_4} are

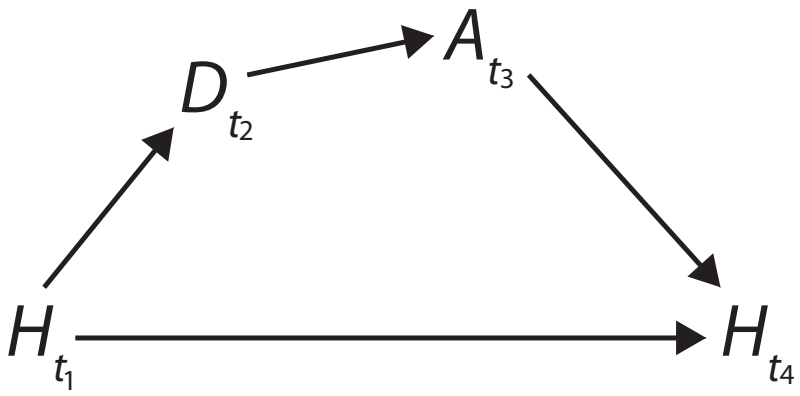

Figure 5.1: A typical causal structure for the fever-and-aspirin case.

binary variables representing whether some individual, call her Mary, has a fever at times t_1

and t_4, respectively.[2] Both take value 1 when Mary has a fever and 0 when she does not. D_{t_2} is a binary variable taking value 1 when Mary decides at t_2 to take aspirin, and 0 when she decides at t_2 not to take aspirin. A_{t_3} is a binary variable taking value 1 when Mary takes aspirin at t_3, and 0 when she does not take aspirin at t_3. Let me also stipulate that the t_1-t_2 and t_2-t_3 intervals are small ($<$ 5 minutes) and that the t_3-t_4 interval is larger ($\approx$ 60 minutes).

The arrows in Figure 5.1 represent causal relations holding between these variables. There is an arrow between H_{t_1} and D_{t_2} because whether Mary has a fever at t_1 is causally relevant to whether she decides to take aspirin at t_2. There is an arrow between D_{t_2} and A_{t_3} because whether Mary decides to take aspirin at t_2 is causally relevant to whether she actually takes aspirin at t_3. And there is an arrow between A_{t_3} and H_{t_4} because whether Mary takes aspirin at t_3 is causally relevant to whether she has a fever at t_4. In addition to the series of arrows connecting H_{t_1} and H_{t_4} via D_{t_2} and A_{t_3}, Figure 5.1 also includes an arrow from H_{t_1} directly into H_{t_4}. This arrow is there to represent the fact that, in addition to its indirect effect on H_{t_4} via D_{t_2} and A_{t_3}, H_{t_1} also has a direct effect on H_{t_4}: Considered in isolation from its effect on whether she takes aspirin, Mary having a fever at t_1 makes her more likely to have a fever at t_4, an hour later.

Readers may have doubts regarding the causal nature of the direct relationship between H_{t_1} and H_{t_4}. From an interventionist perspective, however, these doubts are unwarranted. One can easily show that H_{t_1} is a direct cause—in the sense of (Woodward, 2003b, 59)—of H_{t_4} relative to the set of variables $\mathbf{V}$: $\{H_{t_1}, D_{t_2}, A_{t_3}, H_{t_4}\}$. It is presumably true that if one were to intervene so as to set the value of H_{t_1} to 1 (i.e. induce a fever in Mary at t_1) and hold A_{t_3} fixed to value 0 (i.e. prevent Mary from taking aspirin at t_3), then the probability distribution over values of H_{t_4} would change, since the probability that Mary has a fever at t_4 (i.e. that $H_{t_4} = 1$) would presumably increase.[3] And the truth of this counterfactual is,

[2] How one precisely defines 'fever' is without consequences for the argument developed below.

[3] Note that since one here holds A_{t_3} fixed to value 0 (i.e. prevents Mary from taking aspirin at t_3), the value of D_{t_2} (i.e. whether Mary decides to take aspirin at t_2) becomes irrelevant.

according to Woodward's interventionist account, sufficient for it to be the case that H_{t_1} is a direct cause of H_{t_4} relative to $\mathbf{V}$, and so sufficient for the presence of an arrow from H_{t_1} directly into H_{t_4} in Figure 5.1 to be warranted from an interventionist point of view. Note that I here talk about changes in the probability distribution over values of H_{t_4}, rather than about changes in the value of H_{t_4}, because the direct relationship between whether Mary has a fever at t_1 and whether she has a fever at t_4 is indeterministic, at least relative to the variables in $\mathbf{V}$: Mary having a fever at t_1 does not guarantee that she will have a fever at t_4, it merely makes it more likely. Woodward's interventionist account is designed to apply both to the deterministic and to the indeterministic case.

5.3 The argument

How is knowledge of the fact that, for individuals like Mary, aspirin relieves fevers relevant to predicting the effect that taking aspirin at t_3 would have on whether Mary had a fever at t_4? For interventionists, to know that aspirin relieves fevers in individuals like Mary is, *by definition*, to know that if one were to intervene on A_{t_3} with respect to H_{t_4}, then the probability of $H_{t_4} = 1$ would decrease.[4] The problem here is that Mary's decision, whatever it turns out to be, cannot be an intervention on A_{t_3} with respect to H_{t_4}, for reasons I will detail below. As a result, knowing how the probability distribution over values of H_{t_4} would change under an intervention on A_{t_3} with respect to H_{t_4} is useless for predicting what would happen were Mary to take aspirin as a result of her decision to do so. What Mary needs to know is what would happen were she to take aspirin as a result of a manipulation that does not qualify as an intervention. And this is precisely what the interventionist account, given the way it understands causal claims, cannot tell her.

Let me explain why Mary's decision cannot be an intervention on A_{t_3} with respect

[4] As in the case of the relationship between H_{t_1} and H_{t_4}, the relationship between A_{t_3} and H_{t_4} presumably is indeterministic relative to $\mathbf{V}$, since taking aspirin at t_3 does not guarantee that Mary will not have a fever at t_4.

to H_{t_4}. The event of Mary deciding to take aspirin at t_2 is the event of D_{t_2} taking value 1, commonly abbreviated $D_{t_2} = 1$. Likewise, the event of Mary deciding not to take aspirin at t_2 is the event $D_{t_2} = 0$. A necessary condition for either event to be an intervention on A_{t_3} with respect to H_{t_4} is that D_{t_2} be an intervention *variable* on A_{t_3} with respect to H_{t_4}. In order for this to be the case, D_{t_2} must satisfy the following four conditions (adapted from Woodward 2003b, 98):

> I1. D_{t_2} causes A_{t_3}.
>
> I2. D_{t_2} acts as a switch for all the other variables that cause A_{t_3}. That is, certain values of D_{t_2} are such that when D_{t_2} attains those values, A_{t_3} ceases to depend on the values of other variables that cause A_{t_3} and instead depends only on the value taken by D_{t_2}.
>
> I3. Any directed path from D_{t_2} to H_{t_4} goes through A_{t_3}...[5]
>
> I4. D_{t_2} is statistically independent of any variable Z that causes H_{t_4} via a directed path that does not go through A_{t_3}.

Focus on condition I4. As one can readily tell from Figure 5.1, the only variable that causes H_{t_4} via a directed path that does not go through A_{t_3} is H_{t_1}. For D_{t_2} to satisfy I4, then, it must be statistically independent from H_{t_1}. But we know that this is not the case, since Mary is more likely to decide to take aspirin when she has a fever than when she does not, i.e. since it is more likely that $D_{t_2} = 1$ when $H_{t_1} = 1$ than when $H_{t_1} = 0$. Variable D_{t_2} thus violates I4. As a result, it is not an intervention variable on A_{t_3} with respect to H_{t_4}, which implies that neither $D_{t_2} = 1$ nor $D_{t_2} = 0$ can be interventions on A_{t_3} with respect to H_{t_4}.[6]

It seems obvious that knowledge of the fact that aspirin relieves fevers should help Mary predict what would happen were she to take aspirin as a result of her decision to do so and should therefore guide this decision. Given the way the interventionist account

[5] A directed path in a graph is, briefly, a sequence of variables $X_1, X_2, \ldots, X_{n-1}, X_n$ such that there an arrow from X_1 into X_2, from X_2 into... X_{n-1}, and from X_{n-1} in to X_n.

[6] Note that D_{t_2} is not a 'soft' or parametric intervention variable—in the sense of (Eberhardt and Scheines, 2007)—either. On Eberhardt and Scheines's view (2007, 986), D_{t_2} must be exogenous relative to $\mathbf{V}$, i.e. have no cause in $\mathbf{V}$, in order to be a parametric intervention variable on A_{t_3} with respect to H_{t_4} in $\mathbf{V}$ (their definition relativizes the notion of parametric intervention variable to sets of variables). Since H_{t_1} is a cause of D_{t_2} and since $H_{t_1} \in \mathbf{V}$, however, D_{t_2} is not exogenous relative to $\mathbf{V}$. See (Eberhardt and Scheines, 2007, 984–987) for the difference between structural, i.e. Woodward-style, and parametric interventions.

understands causal claims, however, it cannot explain why this should be the case. Because Mary's decision whether to take aspirin cannot be an intervention on whether she does take aspirin, the interventionist account has nothing to say about what would happen to her fever were Mary to take aspirin as a result of this decision.

5.4 Objections

Let me briefly address three objections one might advance against the argument raised above.

Objection 1 First, one might object that if one grants interventionist the claim that A_{t_3} is a direct cause of H_{t_4} then, by definition, there must be an intervention variable I such that, for some value i of I, $I = i$ is an intervention on A_{t_3} with respect to H_{t_4} that is temporally followed by a change in the value of H_{t_4}.[7] And if this is so, then the claim that A_{t_3} is a direct cause of H_{t_4} does entail some counterfactual about what would happen to the value of H_{t_4} were the value of A_{t_3} to be changed by an intervention. This is both true and beside the point. Whatever I represents, it cannot be Mary's decision.[8] This means that, whatever its precise content, the counterfactual entailed by the claim that A_{t_3} is a direct cause of H_{t_4}—as interventionists understand this claim—simply is not the one Mary needs to know in order to guide her decision whether or not to take aspirin. And the interventionist account has nothing to say about counterfactuals the antecedents of which describe changes in the values of variables that do not result from interventions.

[7] The fact that no such variable is represented in Figure 5.1 is without consequence here. Woodward does not require that, in order for X_1 to be a cause (direct or otherwise) of X_2 relative to a set of variable $\mathbf{V}$: $\{X_1, X_2, \ldots, X_n\}$, some intervention variable on X_1 with respect to X_2 must be included in $\mathbf{V}$.

[8] In fact, because Woodward (2003b, 132) only requires that interventions be "logically or conceptually possible", $I = i$ need not represent a nomologically possible (a fortiori, actual) event.

Objection 2 Second, one might object that what Mary needs to know is not that A_{t_3} is a cause of H_{t_4} but, instead, that the correlation between these two variables will remain intact if she takes aspirin at t_3. And since the interventionist account, I have assumed, yields the verdict that A_{t_3} is a direct cause of H_{t_4}, it also implies that the correlation between these two variables will remain intact if Mary takes aspirin at t_3. As interventionists understand it, however, the claim that A_{t_3} is a cause of H_{t_4} implies not that the correlation between A_{t_3} and H_{t_4} will remain intact under any changes in the value of A_{t_3} whatsoever but only that it will remain intact under changes in the value of A_{t_3} *that are the results of interventions*. One must be careful not to slip from the second of these claims to the first. Interventionists cannot simply avail themselves of properties typically associated with causal relations (e.g. giving rise to correlations that are robust under changes which, like Mary's aspirin taking, are not the result of interventions) when their own account fails to imply that causal relations have the properties in questions. Again, as with Objection 1, the interventionist account has nothing to say about what happens to correlations between variables when the value of one of these variables is changed in a way other than by an intervention.

Objection 3 Third, one might argue that the problem I have raised for the interventionist account is an artifact of the way I chose to represent the fever-and-aspirin case above, an artifact that will disappear when one modifies this representation. Here are two obvious ways to modify this representation: Add variables (and arrows) or remove variables (and arrows). Let me explain why neither kind of modification will help.

The issue with the first kind of modification—addition—is straightforward. D_{t_2} violates conditions I4 because of the $D_{t_2} \leftarrow H_{t_1} \rightarrow H_{t_4}$ path. Since adding variables and arrows to the graph in Figure 5.1 will not remove this path, doing so will not help avoid the result that D_{t_2} is not an intervention variable on A_{t_3} with respect to H_{t_4}.

The issue with the second kind of modification—subtraction—is almost as straightforward. Suppose one decides to refrain from explicitly modeling Mary's decision and so

removes D_{t_2} from the graph in Figure 1. You might think, after all, that it was a mistake to explicitly represent this decision and that this mistake is the source of the problem I'm arguing interventionists face. And you might think that what Mary needs to know is what would happen were she to take aspirin at t_3 as a result of her having a fever at t_1. H_{t_1}, however, is a direct cause of H_{t_4}, which means that some directed path from H_{t_1} to H_{t_4} does not go through A_{t_3}. H_{t_1} therefore violates condition I3 and thus cannot be an intervention variable on A_{t_3} with respect to H_{t_4} either.[9] Refraining from modeling Mary's decision thus cannot help interventionists. As they understand it, the claim that A_{t_3} is a cause of H_{t_4} does not imply any counterfactuals regarding what would happen were Mary to take aspirin at t_3 as a result of her having a fever at t_1.

5.5 A contrast: Cartwright's probabilistic account

One might think that, in asking the interventionist account for a bridge connecting causal claims to predictions regarding the effects of our actions, I am asking it for more than any account of causation can provide. But I am really only asking interventionists for something that (1) they claim to provide and (2) other accounts of causation, e.g. the probabilistic account developed in (Cartwright, 1979), do provide.

According to Cartwright's probabilistic account of causation, aspirin relieves fevers iff:

$$P(\text{fever relief}|\text{Aspirin} \wedge K_i) > P(\text{fever relief}|\neg\text{Aspirin} \wedge K_i) \text{ for every } K_i \qquad (1)$$

Each K_i is a state-description—in the sense of (Carnap, 1946, 50)—over the set $\{C_j\}$, where each C_j is a cause of fever relief, subject to the restrictions discussed by Cartwright (1979, 423). For instance, $\{C_j\}$ must contain neither taking aspirin itself nor any of the causes of

[9] One might think that the issue is with condition I4. See (Woodward, 2003b, 100–102) for the rationale behind the inclusion of I4 as one of the conditions a variable must satisfy to be an intervention variables. To be brief, removing I4 from this set of conditions is not a viable option for interventionists.

fever relief that are effects of taking aspirin. To put it briefly, each K_i gives an exhaustive description of the causes of fever relief (Are they present? Are they absent?) other than taking aspirin and its effects. The requirement that the probabilistic inequality in (1) hold for all K_i is what John Dupré (1984, 170) calls the "requirement of contextual unanimity".

Call m the situation Mary is in and K_m the state-description over $\{C_j\}$ corresponding to this situation. As Cartwright defines the notion, taking aspirin is an *effective strategy* for relieving fever in situation m iff:

$$P(\text{fever relief}|\text{Aspirin} \wedge K_m) > P(\text{fever relief}|\neg\text{Aspirin} \wedge K_m) \tag{2}$$

It should be clear that the truth of (1) entails that of (2).[10] In other words, Cartwright's account implies that if aspirin is a cause of fever relief, then taking aspirin is an effective strategy for relieving Mary's fever. Moreover, and this is the key point here, this implication holds regardless of whether the event bringing about Mary's aspirin taking—e.g. her decision to take aspirin—meets Woodward's conditions for being an intervention. Note, for instance, that Mary's decision need not be uncorrelated with causes of fever relief other than taking aspirin (e.g. with whether or not she has a fever before making her decision) in order for inequality (2) to hold and so for aspirin taking to be an effective strategy for relieving her fever, by contrast with what Woodward requires in order for D_{t_2} to be an intervention variable on A_{t_3} with respect to H_{t_4}.

Cartwright's account thus provides a straightforward explanation for the practical usefulness of causal knowledge in the fever-and-aspirin case. Since the causal law according to which aspirin relieves fevers entails that taking aspirin is an effective strategy for relieving

[10] I here assume, for the sake of simplicity, that the probabilistic inequality in (1) holds for all members of the population Mary belongs to, and so for Mary herself. Note that Cartwright (1979, 435) provides an additional principle connecting conditional probabilities to probabilities of counterfactuals, a principle stating that $P(\text{Aspirin} \boxdot\!\!\rightarrow \text{fever relief}|m) = P(\text{fever relief}|\text{Aspirin} \wedge K_m)$. One can thus redefine the notion of effective strategy in the following way: Taking aspirin is an *effective strategy* for relieving fever in situation m iff $P(\text{Aspirin} \boxdot\!\!\rightarrow \text{fever relief}|m) > P(\neg\text{Aspirin} \boxdot\!\!\rightarrow \text{fever relief}|m)$. And this inequality is implied by (1) just as much as (2) is.

Mary's fever, it is no wonder that knowing that aspirin relieves fevers should be relevant to guiding her decision. The same claim holds of probabilistic accounts of causation that reject the requirement of contextual unanimity (see e.g. Dupré 1984) so long as they take the truth of (2) to entail that, in situation *m* (i.e. for Mary), taking aspirin relieves fevers.

The point here is not to argue that Cartwright's probabilistic account of causation is correct or that it is superior to Woodward's interventionist account. It is simply to give an example of an account of causation which, unlike Woodward's, does explain the practical usefulness of causal knowledge by providing a bridge leading from causal claims to predictions regarding the effects of our actions. Cartwright's account explains why knowledge of the fact that aspirin relieves fevers should guide Mary's decision when Woodward's cannot.[11] And part of the reason why it succeeds in doing so is that it does not require that Mary's decision be an intervention on her taking aspirin.

5.6 Conclusion: The ramifications

I said in Section 5.2 that the causal structure depicted in Figure 5.1 is typical of cases in which one has a fever and is pondering whether to take aspirin. What I meant by this is that the graph in Figure 5.1 can be used to accurately model the fever-and-aspirin case for many individuals at many times, not just for our hypothetical Mary at some particular time. But there is another, more interesting sense in which the causal structure depicted in Figure 5.1 is typical: It is typical of cases in which the effect of interest (e.g. a medical condition) is what prompts the introduction of the putative cause (e.g. a medical treatment). In most cases of this kind, the variable representing the decision whether or not to treat will fail to be an intervention variable on the treatment with respect to the post-treatment condition. This is because, as in the aspirin-and-fever case, the decision to treat and the post-treatment

[11] As should be obvious, the claim that aspirin relieves fevers, as interventionists understand it, implies neither (1) nor, a fortiori, (2). Interventionists thus cannot ride piggy-back on the explanation for the practical usefulness of causal knowledge offered by Cartwright.

condition will be related as effects of a common cause, namely the pre-treatment condition.

This is true for medical cases, but also for cases in which the 'treatment' is a public policy. Consider Project STAR, the class-size reduction policy implemented in 1985 in Tennessee with the aim of improving the test scores of elementary school students. The decision by then-governor Lamar Alexander to implement this policy was not an intervention on class sizes with respect to the post-implementation test scores. This is because his decision was not statistically independent from pre-implementation test scores—since his decision was caused by these scores being lower than desired—and because pre-implementation scores have a direct effect (i.e. an effect not mediated by Alexander's decision and the implementation of the class-size reduction policy) on post-implementation scores, just as whether Mary has a fever at t_1 has a direct effect (i.e. an effect not mediated by D_{t_2} and A_{t_3}) on whether she has a fever at t_4. What this means is that the interventionist account cannot explain the relevance of the causal claim 'In some circumstances, smaller classes lead to an improvement in test scores' to predicting what would happen in Tennessee upon implementation of Project STAR (where test scores did in fact improve as a result of the reduction in class sizes). And it therefore also cannot explain why knowledge of this claim should have guided Alexander's policy decision.

It is because cases of this kind are ubiquitous—in addition to being of great importance—that I claimed, as the end of Section 5.1, that the interventionist explanation for the practical usefulness of causal knowledge does not apply to a great many cases. There may, of course, be cases in which the causal structures in which agents are embedded are such that the decisions they make are genuine interventions. In such cases, the interventionist story regarding the practical usefulness of causal knowledge will hold true, assuming the interventionist analysis of causation to be correct. But these cases are, by all accounts, few and far between. I must conclude, then, that despite what its advocates claim, the interventionist account does not explain the practical usefulness of causal knowledge (except,

maybe, in a few select cases). It thus falls short of the standards set by the functionalist approach Woodward advocates.[12]

[12]The argument developed in this paper echoes that developed in (Reiss and Cartwright, 2005) regarding Pearl's 'structural model' semantics for counterfactuals (Pearl, 2000, §7.1). Because this semantics calls for the antecedents of counterfactuals to be realized by what are, roughly, Woodward-style interventions, it is of very limited use for evaluating 'real world' policy counterfactuals, i.e. counterfactuals regarding what would happen were one to implement a particular policy not as a result of an intervention but as a result of a process (e.g. a decision process) that does not qualify as an intervention.

Chapter 6

Is Race a Cause?

6.1 Introduction

Scientists in many disciplines (economics, epidemiology, etc.) routinely treat race as a cause. Economists who study labor market discrimination, for instance, often build models involving race as an independent variable and interpret estimates of the coefficient attached to it as estimates of the causal effect of race. This practice conflicts with the view held by leading advocates of the counterfactual approach to causal inference (henceforth 'CFA') who argue that, since race is a necessary property of individuals, one cannot coherently treat it as a cause.

Important issues hang on the outcome of this debate between practitioners and theorists of causal inference. If race is not a cause, then the coefficients attached to variables representing race cannot represent the causal effect of race. But then what, if anything, do they represent? And if these coefficients cannot represent the causal effect of race, then is it legitimate to use data on race to estimate them? Should studies that purport to measure the causal effect of race (e.g. on earnings or on access to health care) be funded? And should social and health policies be based on results from such studies?

After a brief introduction to the CFA (Section 6.2), I present the argument against

race being a cause (Section 6.3). I then raise objections against two of its premises (Section 6.4) and sketch a positive argument for race being a cause (Section 6.5).

6.2 The Counterfactual Approach

The CFA, developed primarily by Rubin (see e.g. Rubin 1974), is the dominant approach to causal inference in statistics and in many social and biomedical sciences. It has roots in the work of Fisher and Neyman on agricultural experiments.

When only one cause is considered, counterfactual causal models essentially have the following components:

- A population of units $i \in U$
- A binary causal exposure variable D taking value $d_i = 1$ when i is exposed to the cause (is in the 'treatment' state) and $d_i = 0$ when i is not (is in the 'control' state).
- Two potential outcome variables Y^1 and Y^0, where y_i^1 represents the value of the effect for i when i is exposed to the cause and y_i^0, the value of the effect for i when i is not exposed to the cause.

The individual-level causal effect (ICE) of D for i is typically defined as follows:

$$\delta_i = y_i^1 - y_i^0$$

The ICE is equal to the difference between the value of the effect when i is exposed to the cause and the value of the effect when i is not. Since a given unit cannot be both exposed to the cause and not exposed to it at once, only one of y_i^1 and y_i^0 can be observed for any unit. If i is exposed to the cause, the value of y_i^1 is observable while the value of y_i^0 is counterfactual: It is the value the effect *would* have taken had i not been exposed to the cause, hence the

name of the approach. Because only one of y_i^1 and y_i^0 can be observed, δ_i cannot be observed. Holland dubs this the "fundamental problem of causal inference" (1986, 947).

There are various solutions to this problem, both in experimental and in observational contexts. These solutions provide techniques for estimating the ICE and other causal effects one can build from it. My concern here is not with the problems that race might raise for the application of these estimation techniques. It is, rather, with the problems that race allegedly raises for the very definition of causal effects, and of the ICE in particular.

6.3 The Argument Against Race Being a Cause

The argument developed by advocates of the CFA against race being a cause can be reconstructed as follows:

1. Race is a necessary property of units.

2. If i is of race r, then it is impossible for i to have been of another race r'. (from 1).

3. Counterfactuals of the form 'Had i been of race r' instead of r, then…' cannot be (non-vacuously) true. (from 2).

4. The ICE of race is undefined. (from 3 and the definition of ICE).

5. For all x, if x is a cause, then its ICE is defined.

$\therefore$ Race is not a cause. (from 4 and 5).[1]

[1]Note that the argument, thus reconstructed, is immune to the objection raised by Glymour (1986) against Holland (1986). Glymour objects that, "If counterparts [in Lewis's sense] are conceivable–and why not?–then counterfactuals that violate identity conditions are intelligible, and if counterfactuals are intelligible, then causal relations are as well." (1986, 966) If the problem with race is that it is a necessary property of individuals, however, then whether one favors transworld identity–as Holland implicitly does–or counterpart theory is irrelevant, and no appeal to the latter will help. If being of race r is a necessary property of i, then all the counterparts of i also are of race r, and so counterfactuals of the form 'Had i been of race r' instead of r, then…' have impossible antecedents and cannot be (non-vacuously) true.

Let me illustrate this argument. Assume that there are only two races, that D represents race, and that $d_i = 1$ when i is White and $d_i = 0$ when i is Black. Leading advocates of the CFA, such as Rubin and Holland, hold that race is a necessary property, "immutable characteristic" (Greiner and Rubin, 2011), or "attribute" (Holland, 1986, 955) of units. To say that race is a necessary property of units is to say that if $d_i = 1$ (resp. 0), then it could not have been the case that $d_i = 0$ (resp. 1). Because this is so, counterfactuals of the form 'Had it been the case that $d_i = 0$ instead of $d_i = 1$, then the value of Y^0 for i would have been y_i^0' cannot be non-vacuously true when $d_i = 1$ (and conversely when $d_i = 0$). In Holland's words, "attributes of units [e.g. race] are not the types of variables that lend themselves to *plausible states* of counterfactuality." (2003, 14, emphasis original)[2] Because no such counterfactual can be non-vacuously true, however, the ICE of race is undefined, and this regardless of what effect the potential outcome variables Y^1 and Y^0 represent (earnings, education, etc.).[3] And since the ICE of race is undefined, race is not a cause.

The consequences of this view are important. If race is not a cause then, as Greiner and Rubin point out, "attempts to infer the causal effects of such traits [as race] are incoherent." (2011, 775) Holland goes further by claiming that, "Attributing cause to RACE is merely confusing and unhelpful" and that, "Obscuring [the topics of discrimination and bias] with simplistic calculations that do not attend to the proper role of RACE in a causal study helps no one." (2003, 24)

So, do the many scientists who treat race as a cause waste time and resources on incoherent studies that only obscure important topics like racial discrimination? I do not believe so and now turn to two objections to the argument against race being a cause.

[2]Holland adds: "Because I am a White person, it would be close to ridiculous to ask what would have happened to me had I been Black." (2003, 14)

[3]The same point applies mutatis mutandis to other causal effects defined in the CFA, e.g. the average causal effect defined over U as $E[Y^1] - E[Y^0]$.

6.4 Against the Argument Against Race Being a Cause

6.4.1 Why Believe Premise 5?

According to premise 5 in the argument against race being a cause, having a well-defined ICE is a necessary condition for being a cause. To believe this premise is to believe that every cause can be handled by the CFA. There are good reasons, however, to doubt this claim.

Consider, for instance, the case of primary school performance: According to Holland himself, scholastic achievement in primary school cannot be treated as a cause of the choice of secondary school by the CFA because its ICE is undefined (1986, 955).[4] Assuming for a minute that Holland is correct in his assessment, the right conclusion to draw here does not seem to be that scholastic achievement is not a cause of school choice. This is so because there are very good reasons to think that how well a student does in primary school has a causal effect on what secondary school she chooses to attend, e.g. by determining what schools she is admitted to. The right conclusion to draw, rather, seems to be that some genuine causes cannot be handled by the CFA, and therefore that having a well-defined ICE is not necessary to be a cause.[5]

This conclusion is bolstered by the existence of frameworks for causal inference, e.g. Ragin's qualitative comparative analysis framework (1987), that do not rely on counterfactuals to define causal effects and which can thus treat properties whose ICE is undefined as

[4]Holland holds this view because he thinks that, "It is difficult to conceive of how scholastic achievement could be a treatment in an experiment…" (1986, 955) and because, as a result, he thinks that scholastic achievement, like race, does not lend itself to "plausible states of counterfactuality". Though Holland's reasoning is faulty–because it relies on a principle that advocates of the CFA should reject, as I will argue in §4.2–let me assume here, for the sake of argument, that the conclusion he reaches is true.

[5]The same conclusion seems warranted in the case of the ICE of the age at which a student starts school on her first grade test scores, a causal effect econometricians Angrist and Pischke dismiss as "impossible to interpret" (2009, 7) in the CFA and as giving rise to "a fundamentally unidentified question" (op. cit., 5). Because there are good reasons to think that the age at which a student starts school has a causal effect on her first grade test scores, the fact that this causal effect is "impossible to interpret" in the CFA suggests that there are genuine causes the CFA cannot handle.

causes.

6.4.2 Why Believe Premise 1?

Why should one believe the claim that race is a necessary property, or attribute (in Holland's terms), of units? How do advocates of the CFA justify this claim? Their justification derives entirely from an application of what I will call 'Holland's rule' (or 'HR'). As Holland originally formulates it, HR states that,

> If the variable *could be* a treatment in an experiment (even one that might be impossible to actually pull off due to ethical or practical issues), then the variable is [...] correctly called a *causal variable*. (Holland, 2003, 9, emphasis original)

It is important to note that, for Holland, attributes and causal variables form a partition of the set of properties of a unit: A property is an attribute if and only if it is not a causal variable.[6] Holland claims that race could not be a treatment in an experiment and, applying HR, concludes that it is not a causal variable but, rather, an attribute or necessary property (ibid.). As should be obvious, Holland's argument is fallacious given the way he formulates HR: It denies the antecedent of HR and infers the negation of its consequent. I will here adopt a charitable reading according to which being a treatment in some possible experiment is both sufficient and *necessary* for a property to be a causal variable. The proper formulation of HR–and the one I will discuss below–is thus as follows:

(HR) A property is a causal variable if and only if it could be a treatment in an experiment (even one that might be impossible to actually pull off due to ethical or practical issues).

Greiner and Rubin agree with Holland's line of argument and invoke "the impossibility of manipulating such traits [as race] in a way analogous to administering a treatment in

[6]Note that I use 'property' and 'variable' as synonyms in this paper (as does Holland in his writings). This is without consequences for the arguments developed.

a randomized experiment" (2011, 775) as the main source of the incoherence of studies purporting to estimate the causal effect of race.

Given the biconditional formulation of HR given above, Holland's argument to the effect that race is an attribute is valid. There are, however, two issues with HR. First, it is the wrong rule for advocates of the CFA to follow, i.e. advocates of the CFA should see HR as false. According to the CFA, for the ICE of *D* on *i* to be defined, there must be some counterfactual state in which *i* is not exposed to *D*, assuming that *i* actually is exposed to *D*. In other words, it must be possible for *i* not to have been exposed to *D*. But why think that the possibility of such a state requires the possibility of an experiment resulting in it being the case that *i* is not exposed to *D*? To hold this view is to hold the implausible view that it is possible that *p*–where *p* is of the form '*i* is exposed (resp. not exposed) to *D*'–only if it is possible for there to be an experiment resulting in it being the case that *p*. The right slogan for the CFA thus is not "No causation without [some possible experimental] manipulation" (Holland, 1986, 959) but, rather, 'No causation without counterfactual states'. This slogan is less catchy but more faithful to the way the CFA defines causal effects (e.g. the ICE).

One might object that HR was intended by Holland not as a strict rule but as a heuristic. It is true that Holland prefaces his discussion of HR by saying that, "There is no cut-and-dried rule for deciding which variables in a study are causal and which are not." (2003, 9) But note that, despite this caveat, he *does* apply HR as a "cut-and-dried" rule, since he takes the supposed violation of HR by race to be sufficient to establish the conclusion that race is an attribute and so is not a causal variable (op. cit., 10). It should also be noted that HR fares no better as a heuristic than it does as a strict rule. I have claimed above that the possibility of an experiment resulting in *i* not being exposed to *D* is not necessary for it to be possible that *i* is not exposed to *D*. If so, however, then there is no reason to take the inconceivability of such an experiment to be a reliable guide to the impossibility of a state in which *i* is not exposed to *D*.

The second issue with HR is that it is vague and that, as a result, it is unclear that it is genuinely impossible for there to be an experiment in which race is the treatment. Consider the following hypothetical (randomized) experiment: Assume that the race r_i of unit i is a function $r_i = f(b_i, e_i)$ of biological (b_i) and environmental (including social and cultural) factors (e_i).[7] Imagine that values of b_i and e_i, and thus also of r_i, are randomly assigned to embryos 30 days after conception. The biological factors are assigned via genetic engineering and the environmental factors are assigned by swapping embryos between mothers.

This experiment has not been carried out, is morally objectionable, and is (presumably) practically impossible given present science and technology. But, according to Holland himself, this does not mean that this experiment is impossible. HR, however, does not give one any more guidance regarding what it means for an experiment to be possible. I take it to be obvious that this experiment is logically possible. This experiment also seems to be nomologically possible, i.e. it does not seem that carrying it out would require the violation of any laws of nature. Is this experiment also conceptually possible? Not if your favorite concept of race implies that values of b_i and e_i, i.e. biological and environmental factors, are not sufficient to determine an individual's race.[8] But if your favorite concept of race has this implication, then why think that it is the right concept for economists or epidemiologists studying race to be using? An argument is needed here to justify the claim that these scientists should work with such a concept of race.

There are thus good reasons to think that the experiment described above is logically, nomologically and conceptually possible, and so good reasons to think that it is possible for race to be a treatment in an experiment, even a randomized experiment. It might be, of course,

[7] You can set the relative weights of b_i and e_i however you like. This set-up is intended to be as neutral as possible between concepts of race.

[8] This will be the case, for instance, if you think that geneological factors (e.g. the identity of i's biological parents) contribute to determining i's race (and are not screened off by values of b_i). Thus, if you think that races are biological groups unified by Section 6. Genealogical relations (see e.g. Hardimon 2012), then you should think that what the experiment described above randomly assigns is not genuinely race.

that the relevant notion of possibility is neither logical nor conceptual nor nomological possibility. And it might be that the concept of race economists and epidemiologists–among others–ought to adopt is in fact one which implies that values of b_i and e_i are not sufficient to determine i's race. It should be clear, however, that one must commit to rather specific views of race and of the notion of possibility at work in HR–and have a good justification for these commitments–in order to defend the view that race violates this rule.

In brief, then, even if HR was the right rule for advocates of the CFA to follow, a view I have argued against, it is doubtful that its application would yield the conclusion that race is an attribute or necessary property.

6.5 A Positive Argument for Race Being a Cause

Consider an imaginary society in which there are two exclusive and exhaustive racial groups, A and B. Assume that there is a wage gap between *As* and *Bs* in this society: *As* receive wages that are uniformly 30% lower than the wages received by *Bs* occupying equivalent jobs. Assume, further, that all the units in the population, be they A or B, are perfectly homogeneous regarding the causes of wages (other than, possibly, race), e.g. they received the same degree from the same school, they have the same work experience, they have the same interpersonal skills, they are equally productive, they have the same preferences regarding wages, etc. Assume, finally, that there is only one employer in this society, and that this employer fixes the wages of every worker.

What is the mechanism generating the wage gap in this society? What explains the fact that some A worker, call her w_A, receives wages 30% lower than those of a B worker, call her w_B, occupying an equivalent job? One straightforward answer is that w_A receives wages 30% lower than those of w_B because she is an A and because the employer believes the work of *As* to be worth 30% less than that of *Bs*. In other words, the fact that w_A is an A, together with the employer's belief about the relative worth of the work of *As*, is the cause

of her receiving wages 30% lower than those of w_B. And, given the set-up described in the previous paragraph, it seems intuitively correct to say that, had w_A been a B instead of an A, she would have received higher wages.

This commonsensical explanation is a causal explanation, since it purports to explain the wage gap by citing its causes, and one of the causes it invokes is the race of w_A and of other A workers. This explanation thus is unavailable to those holding the view that race is not a cause. Indeed, according to advocates of the CFA, counterfactuals about what the wages of w_A would have been like had she been a B instead of an A have impossible antecedents. But what might then explain the wage gap between *As* and *Bs*? I examine the most prominent alternative explanation below.

According to the view defended by Greiner and Rubin (2011), among many others, races themselves play no causal role in generating the wage gap between *As* and *Bs*. What causally explains this gap, rather, are *perceptions* of race. More precisely, what explains the fact that w_A receives wages 30% lower is not her race in combination with the employer's belief regarding the relative worth of the work of *As*, but the perception of her race by the employer in combination with this same belief. According to this view, then, coefficients attached to variables representing race in models should be understood as representing the causal effect of perceptions of race rather than the causal effect of race itself. There are several problems with this alternative explanation, however. I examine three below.

First, if the move to perceptions is warranted in the case of race, then why shouldn't it be warranted for other properties of units as well? Why not think that, rather than work experience (or education, or…), it is the *perception* of work experience (or education, or…) that is causally relevant to an individual's wages, for instance? The move from race to perceptions of race seems rather ad hoc and, in the case of Greiner and Rubin at least, is largely motivated by the assumption that race is not a cause, an assumption which, I argued in §4, is unjustified.

Second, in the imaginary society I described, it is easy enough to determine who's perception it is that is causally relevant to explaining the wage gap, since there is only one employer. But what if there were many employers, and what if the wages of *As* were on average, rather than uniformly, 30% lower than those of *Bs*? Who's perception would then be causally relevant? The collective perception of all the employers? Or the collective perception of only those employers who believe the work of *As* to be worth less than that of *Bs*? If one is to appeal to perceptions of race to explain any real wage gap between racial groups, then one needs answers to these questions. Greiner and Rubin themselves point out the difficulty in answering these questions as one limitation of this approach (ibid., 783-84). And the problem is more severe even when one considers studies of the effect of race on education or access to health care: What is the proper interpretation in terms of perceptions of race of the causal effects estimated by these studies? The move from race to perceptions of race thus raises as many questions as it answers.

Third, what is it that causes the employer in the imaginary society I described to perceive *A* workers, e.g. w_A, to be *As*? If race is not a cause, then what causes the employer to perceive w_A to be an *A* cannot be the fact that she is an *A*, i.e. it cannot be her race. The most plausible alternative here seems to be to claim that what causes the employer to perceive w_A to be an *A* is the instantiation by w_A of a set of features F the presence of which is strongly correlated with, but does not constitute, being a *A*. Consider the case in which $F : \{\text{skin color S}\}$. A question immediately arises: If the employer perceives w_A to be an *A* solely on the basis of her skin color and then proceeds to give her wages 30% lower than *Bs* in equivalent job on the basis of this perception then is this case properly described as a case of racial discrimination? Or is it a case of discrimination on the basis of skin color?

Insofar as the employer de facto equates race and skin color when, by assumption, they are not identical, it seems more appropriate to describe this case as one of discrimination on the basis of skin color than as one of genuinely racial discrimination. Consider the fact

that, if the correlation between being an *A* and being of skin color S is less than perfect, then the employer will discriminate against some non-*As* and fail to discriminate against some *As*. In other words, the line between workers that are discriminated against and workers that are not will cut across racial groups to follow the line between skin colors. The view that this case is not one of racial discrimination is further supported by standard definitions of 'racial discrimination', e.g. the definition formulated by a panel of the US National Research Council and which equates racial discrimination with "*differential treatment on the basis of race* that disadvantages a racial group. . . " (Blank et al., 2004, 39, emphasis original)

In brief, if perceptions of race are not caused by race but, rather, by features the instantiation of which is merely correlated with race, then it is not clear that discrimination on the basis of these perceptions is properly described as racial discrimination.[9] In other words, it is doubtful that Greiner and Rubin can explain cases of genuinely *racial* discrimination without assuming race to be a cause.

The alternative explanation developed by Greiner and Rubin thus does not seem nearly as satisfactory as the commonsensical explanation sketched above, and which assumes race to be a cause. Of course, Greiner and Rubin's approach is not the only possible alternative, and neither has it been fully worked out yet. But it is by far the most prominent in the literature. That it faces significant difficulties thus provides *some* support for the claim that one must assume race to be a cause in order to explain racial discrimination.

6.6 Conclusion

I have defended the view that the argument developed by advocates of the CFA against race being a cause is unsound because two of its premises are false. And I have sketched a positive argument to the effect that race must be assumed to be a cause in order to explain instances of racial discrimination. There thus seems to be good reasons not to

[9] And so it is not clear that these perceptions are properly called 'perceptions of race' in the first place.

follow Holland, Rubin, and other advocates of the CFA in a wholesale dismissal of attempts to draw causal inferences about race as "incoherent" (2011, 775).

I have said little so far about debates in the philosophy of race. If the arguments developed above are sound, then it seems that philosophers of race should ensure that, whatever concept of race they think ought to be used by scientists studying the role of race, their account of this concept implies that race can be a cause.

The debate over the causal status of race examined in this paper also gives a useful example of a case in which philosophers of science can, and should, contribute to clarifying the debate and critically examine the assumption made by the scientists involved. This is what I have tried to do above.

6.7 Acknowledgments

This chapter has been published in *Philosophy of Science* and is reprinted here with permission from the publisher:

- Marcellesi, A. 2013. "Is Race a Cause?" *Philosophy of Science* 80(5): 650-659.

Chapter 7

External Validity: Is There Still a Problem?

7.1 Introduction

It is customary, at least since (Campbell, 1957), to distinguish between two properties of the conclusions drawn from studies purporting to estimate causal relationships: internal validity and external validity. This distinction is typically drawn as follows (see e.g. Cook and Campbell 1979, 37): A causal conclusion drawn from a study is internally valid if and only if it is true of the population on which the study was conducted. And it is externally valid if and only if it is true not just of the study population but also of some other population(s) as well. Though there is fairly wide agreement on the strictures a study must satisfy in order for the causal conclusions one draws from it to be internally valid, matters tend to muddier when it comes to external validity. There is no widespread agreement on the conditions under which one can infer from the truth of a causal claim in a study population to its truth in some other population.

In this paper I propose, first, to distinguish two kinds of external validity inferences, predictive and explanatory. I will then argue that we have the correct answer—developed

independently by Nancy Cartwright and Jeremy Hardie on the one hand and Judea Pearl and Elias Bareinboim on the other—to the question of the conditions under which predictive external validity inferences are good inferences. If this claim is correct, then it has two immediate consequences: First, Daniel Steel's demand that any acceptable account of external validity inferences break what he calls the 'Extrapolator's Circle' is misplaced. Second, some external validity inferences are deductive, in contradiction with the widely accepted view according to which they are inductive.

I will say little about external validity inferences of the explanatory kind. My hope in this paper is to show that, as far as predictive external validity inferences are concerned, the problem has been solved.

7.2 A Classification of External Validity Inferences

Issues of external validity (EV) typically arise in one of the two following contexts. They arise, first, in Evidence-Based Policy (EBP), where EV inferences are often drawn from the results of field experiments—e.g. randomized controlled trials (RCTs)—aiming to estimate the effects of particular policies. Second, issues of EV arise when interpreting the results of laboratory experiments, especially in the cognitive and social sciences, from cognitive psychology to experimental economics. Let me give an example of what EV inferences look like in each of these two contexts, starting with the case of EBP.

Consider the following schematic case, which I will use for illustration throughout the paper. Imagine that Mary is the superintendent of some school district d_1, a district which includes n high schools $S_1, \ldots, S_n$. Imagine also that Mary is dissatisfied with the scores achieved by her high school students on the SAT (Scholastic Assessment Test) and that she would like to take measures to increase these scores. Call Y_i the discrete variable representing the mean SAT score achieved by students from high school S_i. Because Mary is in charge of the entire district, and not of any specific school, the quantity she seeks

to increase is the mean district-level SAT score E[Y_i]. A colleague of Mary's who is the superintendent of some neighboring school district d_0 suggests to her that she implement a particular educational policy P, a policy which has been evidenced—say, by a well-conducted RCT—to increase the mean SAT score in d_0. Call X the binary variable representing whether policy P has been implemented ($X = 1$ when it is, 0 otherwise) and call m (where $m > 0$) the estimated size of the effect of $X = 1$ on E[Y_i] in d_0.

Should Mary go ahead and implement policy P in d_1 on the basis of the evidence that P caused an increase in SAT scores in d_0? Only if the causal claim that implementing P leads to an increase of magnitude m in SAT scores is externally valid, i.e. only if it is true not just in d_0 but also in d_1. The question of EV thus is central to EBP: Any policy decision that is based on evidence gathered in locations (whether spatial or temporal) other than the one targeted by the proposed policy must, it seems, involve some EV inference. And though the case described above is fictional, it is representative of the inferential problem typically faced by practitioners of EBP.

The second context in which issues of EV typically arise is in discussions of results from laboratory experiments. Consider the following classic example from experimental economics. It is often the case that, in order to conduct cost-benefit analyses of environmental policies, economists and policy makers must first assess the monetary value of various nonmarket goods (e.g. clean air). One way to do so is to assess how much members of the population targeted by the policy in question are willing to pay for the good, assuming that they do not already possess it. Another way is to assess how much they are willing to accept as a payment to part with the good in question, assuming they possess it. According to standard preference theory, 'Willingness to Pay' (WTP) and 'Willingness to Accept' (WTA) are two measures of the same quantity, namely of the monetary value the members of some population attach to a nonmarket good.

There is ample experimental evidence, however, that WTP and WTA diverge for

identical goods (see e.g. Starmer 2000): People tend to value the same good differently depending on whether they are asked how much they are willing to pay to acquire it or how much they are willing to accept as a payment to part with it. And they tend to value goods more highly when they possess them than when they do not, i.e. the WTA for a given good tends to by higher than its WTP. This is what has come to be known as the 'endowment effect'. One prominent explanation for the endowment effect is loss aversion (see e.g. Kahneman et al. 1990). According to models incorporating loss aversion, agents are more sensitive to losses than they are to gains, and this is why the amount they are willing to pay to acquire a good they do not possess is lower than what they are willing to accept as a payment to part with the same good. And there is experimental evidence to the effect that loss aversion is in fact a cause of the endowment effect (see e.g. Bateman et al. 1997).

There are at least two ways in which issues of EV arise in this classical case from experimental economics. First, one might wonder whether loss aversion, given experimental evidence to the effect that it is widespread, can help explain past instances in which the endowment effect has been observed, either inside the lab or outside. Second, one might wonder whether, given that one is trying to determine the monetary value of a nonmarket good for a population the members of which are known to be loss averse (with respect to this good), one should expect to observe an endowment effect and so a discrepancy between WTA and WTP (which should in turn inform one's choice of measure). Those are two kinds of EV inferences one might draw from the result of laboratory experiments on loss aversion and the endowment effect.

One might be tempted to classify EV inferences depending on whether they are drawn from the results of field experiments, as in the context of EBP, or from the results of laboratory experiments, as in the context of the economic experiments described above. And, indeed, discussions of EV are often conducted either in one context (see e.g. Cartwright and Hardie 2012 for EBP) or the other (see e.g. Guala 2005 for experimental economics), but

rarely across those two contexts (though see Steel 2008). A more useful classification, I think, is based not on the nature of the experiments from which EV inferences are drawn but on the aim with which they are drawn. And there are two broad aims with which such inferences are drawn: prediction and explanation. Let me formulate the distinction between these two kinds of EV inferences as follows:

- **Predictive EV inference**: An account of predictive EV inference will tell you under what conditions an inference from the claim that the effect of a cause C on E in a population P_1 is of size m to the claim that C will have an effect of size m on E in a distinct population P_2 if it occurs there is a good inference.

- **Explanatory EV inference**: An account of explanatory EV inference will tell you under what conditions an inference from the conjunction of (1) the claim that the effect of a cause C on E in a population P_1 is of size m and (2) the observation of E in a distinct population P_2 to the claim that C has occurred in P_2 and has had an effect of size m on E there is a good inference.

This classification of EV inferences cuts across the two contexts described above. To be sure, most EV inferences drawn from RCTs in the context of EBP are in fact predictive, but they need not be. And while EV inferences drawn from laboratory experiments tend to be explanatory, at least some of them are predictive, as the second EV inference considered above in the case of the effect of loss aversion on the endowment effect illustrates. And the case of the Federal Communications Commission auction for telecommunication licenses—which Guala (2005) discusses under the heading 'Economic Engineering'—is a concrete case in which a predictive EV inference has been drawn from the results of laboratory experiments.

There are various ways one might refine or reformulate the distinction between types of EV inferences drawn above. One could first replace the terms 'predictive' and 'explanatory' by the terms 'prospective' and 'retrospective', since predictive claims typically

are about future events while explanatory claims typically are about past events. One could also appeal to Mill's distinction between the "causes of effects" and the "effects of causes" (Mill, 1843, Book III, Chapter X). Predictive EV inferences are inferences about the effects of causes, i.e. they are inferences the conclusions of which are claims regarding which effects will occur if certain causes occur. Explanatory EV inferences, as I have characterized them above, are by contrast inferences regarding the causes of effects insofar as their conclusions are claims regarding what the causes of observed effects are.

Now that predictive and explanatory EV inferences have been distinguished, let me turn to an argument to the effect that, as far as *predictive* EV inferences are concerned, the problem of EV has been solved.

7.3 The Problem of Predictive External Validity Has Been Solved

The accounts of predictive EV developed by Cartwright and Hardie (2012) on the one hand and Pearl and Bareinboim (2013) on the other share a common core, and this even though they are set in different frameworks. Consider again the educational policy example introduced above. Traditional accounts of predictive EV (e.g. Cook and Campbell 1979) will enjoin Mary to ensure that her school district is sufficiently 'similar', in the relevant ways, to district d_0 before inferring to the claim that implementing policy P in d_1 will cause an increase of size m in the district-level mean SAT score. The Cartwright-Hardie and Pearl-Bareinboim accounts represent important advances because they tell you exactly the ways in which school district d_1 must be 'similar' to d_0 in order for Mary's predictive EV inference to be good.

The insight that is common to both accounts is that causes, including educational policy P, do not produce their effects in a vacuum. The size of the effect of P on SAT scores

in any particular school S_i, for instance, will depend on a variety of school-level factors: The size of the classes in S_i, the quality of the teachers implementing P in S_i, etc. Cartwright and Hardie (2012, 25) call these factors the size of the effect of P on SAT scores depends on "support factors" while Pearl and Bareinboim (2013, 108) call them "difference-generating factors". Let me simply call them 'interactive factors', though I should make it clear that the interaction may be merely statistical, i.e. these factors need not literally interact with the cause of interest.[1]

The idea behind both the Cartwright-Hardie and Pearl-Bareinboim accounts of predictive EV is that the inference from the claim that P had an effect of size m on SAT scores in d_0 to the claim that P will have the same effect if implemented in d_1 is a good inference if and only if the school-level interactive factors for P are 'the same' in d_1 and d_0. Let me briefly explain in more details how the Cartwright-Hardie account develops this idea and how it explicates the meaning of 'the same' in the previous sentence. Limitations of space unfortunately prevent me from here discussing the Pearl-Bareinboim account and comparing it to the Cartwright-Hardie account. Let me simply say that these two accounts converge on similar sets of conditions for predictive EV inferences to be good.

Cartwright and Hardie (2012, 27) follow Mackie in embracing the view that causes are INUS conditions.[2] An INUS condition for some event e is an *Insufficient* but *Necessary* part of an *Unnecessary* but *Sufficient* condition for the occurrence of e. Mackie's classic example is that of a house fire caused by a short circuit (Mackie, 1965). The short circuit is not individually sufficient to produce the fire. Other factors, which I called 'interactive factors' above, are required: The presence of flammable materials, the presence of oxygen, etc. These interactive factors, together with the short circuit, are jointly sufficient for the occurrence of the fire.[3] But note that they need not be necessary. There are other ways to

[1] One could also have considered district-, class-, and even student-level interactive factors here. Let me, however, focus on school-level interactive factors for the sake of simplicity.

[2] They do not, however, hold the view that INUS conditions always are causes, a view which faces well-known counterexamples

[3] It is assumed here that causation is deterministic. One can, however, extend to discussion to the proba-

produce house fires, i.e. there are alternative sets of factors—e.g. sets that have lit cigarettes instead of short circuits—that are sufficient to produce the same fire.

Consider again our educational policy example. In what follows, I will sometimes refer to policy P by using the binary variable representing whether it is implemented, namely X, and to SAT scores by using the variable Y_i. Call $\mathbf{V}$ the vector of variables $\{V_1, \ldots, V_n\}$ each of which represents a school-level interactive factor for the effect of X on Y_i in d_0. If class size is an interactive factor for policy P, for instance, then $\mathbf{V}$ will contain a discrete variable V_i representing the average size of classes in school S_i. A full specification of the configuration of interactive factors for a particular school S_i in d_0 is simply a particular realization $\{v_1, \ldots, v_n\}$ of $\mathbf{V}$. Now define variable Z as ranging over realizations of $\mathbf{V}$, so that each value of z of Z represents a possible configuration of interactive factors for the effect of X on Y_i in a particular school in district d_0. The probability distribution $P_{d_0}(Z)$ simply describes the way in which these interactive factors are distributed in schools in district d_0.

According to Cartwright and Hardie, where predictability can be expected, we can assume that the effect of X on Y_i in a school S_i is governed by a "causal principle" (2012, 23). Without making any commitments as to its precise functional form, one can write out this causal principle as follows:

$$Y_i = f_{S_i}(X, Z) \tag{CP}$$

Let me assume, for the sake of simplicity, that (CP) holds of all schools in d_0. Given this assumption the effect of X on Y_i not at the school-level but now at the district-level for d_0 is simply the sum of the school-level effect of X on Y_i for each value of Z weighted by $P_{d_0}(Z)$, the probability distribution over values of Z in d_0.

Given this set-up, one can formulate the following argument, call it Argument A, to bridge the gap between the effect X had on $\mathrm{E}[Y_i]$ in d_0 and the effect it will have if implemented in d_1:

---

bilistic case by redefining an INUS condition for e as an insufficient but necessary part of an unnecessary but sufficient condition for *an increase in the probability* of e occurring.

1. X had an effect of size m on $\mathrm{E}[Y_i]$ in d_0.

2. The causal principle governing the school-level effect of X on Y_i in d_0 also holds in d_1—i.e. (CP) holds of all schools in d_1.

3. The interactive factors summarized by $\mathbf{V}$ are distributed in the same way in d_1 as in d_0, i.e. $P_{d_0}(Z) = P_{d_1}(Z)$.

$\therefore$ X will have an effect of size m on $\mathrm{E}[Y_i]$ if implemented in d_1.

What Argument A says, to put it slightly differently, is that if the same educational policy P is implemented in both d_0 and in d_1 then, if the interactive factors for P are distributed in the same way in both districts and if they interact with P in the same way in both, then P will produce an effect of size m on the district-level mean SAT score in d_1 just as it did in d_0.[4] This is true as a matter of definition: Because the conjunction of the occurrence of P and of its required interactive factors in d_0 is assumed to be *sufficient* for the production of an effect of size m on $\mathrm{E}[Y_i]$, the occurrence of a qualitatively identical set of conditions in d_1 cannot but produce the same result.

According to Cartwright and Hardie, then, a predictive EV inference is a good inference just in case it takes the form of a sound deductive argument of the kind of Argument A.[5] Their account thus gives a clear and unambiguous answer to the question of the circumstances in which two populations are 'similar' enough that one can extrapolate a causal claim from one to the other.

[4] Assuming here that there are no competing sets of jointly sufficient conditions poised to produce an opposite contribution to $\mathrm{E}[Y_i]$ in either d_0 or d_1.

[5] It should be clear that the existence of such an argument is sufficient but not necessary for the truth of the conclusion that X will have an effect of size m on $\mathrm{E}[Y_i]$ if implemented in d_1. My focus here, however, is on the conditions under which one can infer the truth of this conclusion on the basis of claims about the effect of X in d_0 and about similarities between d_0 and d_1, not on the conditions under which this conclusion is true.

7.4 Two Consequences

Steel's 'Extrapolator's Circle'? Steel (2008, 4) claims that any satisfactory account of EV inferences (or extrapolations, as he calls them) should break what he calls the 'Extrapolator's Circle' (EC). What would that mean in the case I have used above? To break the EC in this case would be to show that one can establish that d_0 and d_1 are sufficiently similar (in the relevant respects) for a predictive EV inference from the former to the latter to be good without having to know so much about d_1 that one could accurately predict the effect P will have on SAT scores if implemented there without appealing to information about the effects of P in d_0. The idea behind the requirement that an account of EV inference break the EC is that such an account should not imply that EV inferences are redundant or pointless.

Does the Cartwright-Hardie account of predictive EV inferences break the EC? It seems that, in order to establish that premises 2 and 3 in Argument A above are true, one must know *a lot* about d_1, since one must know what the relevant causal principle is in d_1 and how the required interactive factors are distributed there. And if one possessed this knowledge, then would not one be in a position to accurately predict the effect of P on SAT scores in d_1 without having to appeal to information about d_0? If this is the case, then the Cartwright-Hardie account does not break the EC and so it is not, by Steel's standards, acceptable as an account of predictive EV inferences.

Should one then reject the Cartwright-Hardie account? This seems too hasty a conclusion. I claimed above that this account gives the correct answer to the question of the circumstances under which predictive EV inferences are good. Assuming that this claim is true, what more could one want of an account of EV inferences? One might argue that such an account should, in addition, tell you how go about drawing such inferences. But one should be careful to distinguish between *analyses* (for lack of a better word) of the conditions under which EV inferences are good inferences and *methods* for drawing such

inferences. Cartwright and Hardie (2012) provide both in their book-length treatment of EV inferences in EBP. For instance, they give methods for finding out about interactive factors. Breaking the EC might be a legitimate desideratum for a method for drawing EV inferences, but it is not for analyses of the conditions under which they are good inferences. Such analyses should be evaluated according to whether they provide clear and principled accounts of the conditions under which one can infer from the truth of a causal claim in one population to its truth in another population. And, by this standard, the Cartwright-Hardie account is successful.

Deductive EV Inferences EV inferences, whether predictive or explanatory, are widely assumed to be inductive. Guala (2005, 196) expresses this received view when he claims, for instance, that "external validity inferences surely involve an inductive step of some sort". If the Cartwright-Hardie account of predictive EV inference is correct, however, then this assumption is false: Some EV inferences are deductive.

One might, to be sure, be uncertain as to the truth-values of the premises required for good EV inferences according to the Cartwright-Hardie account. One might, for instance, be uncertain as to whether the causal principle governing the effect of X on $\mathrm{E}[Y_i]$ is the same in d_0 and d_1. But an uncertainty regarding the truth-values of the premises of Argument A is obviously not the same thing as an uncertainty regarding whether the conclusion of this argument is entailed by its premises. I hope to have shown that, given the way Cartwright and Hardie define the notions of causal principle and interactive factors, there is little room to dispute the claim that Argument A is a valid deductive argument.

One might worry that, by interpreting predictive EV inferences as purported deductive arguments of the kind of Argument A, the Cartwright-Hardie account places the bar too high. Since it seems that, in any concrete case, at least some of the required interactive factors are likely to be 'local' (i.e. be present in the study population but not in the target one), the Cartwright-Hardie account implies that predictive EV inferences will rarely (if

ever) be good inferences. Is this skeptical upshot problematic for Cartwright and Hardie? It is, after all, what motivates Steel, Guala and others to reject an earlier account of EV inference due to Hugh LaFollette and Niall Shanks (1995).

The skeptical upshot of the Cartwright-Hardie account would be problematic if (i) the predictive EV inferences we actually draw often are good inferences or (ii) if it implied that Mary, the superintendent of d_1, cannot learn anything valuable from the effect of P on SAT scores on d_0. The evidence for (i), however, is hardly overwhelming. And (ii) is plainly false.

Let me start with (i). Imagine that Mary implements P in d_1 on the basis of its effect on SAT scores in d_0 and that, as it happens, P produces the same result in d_1 as in d_0. Should we conclude that the inference drawn by Mary was a good predictive EV inference? No, because for all we know, Mary might have used naive induction ('It worked there, therefore it will work here') and simply gotten lucky. In the vast majority of cases, it is impossible to determine whether policy decisions that led to desirable results were in fact based on good predictive EV inferences. This is because, in the vast majority of cases, the process by which evidence is used to arrive at a policy decision is not public. And cases of bad predictive EV inferences leading to policy decisions with undesirable results are not hard to find (see e.g. Cartwright and Hardie 2012, II.B.4). An argument against the Cartwright-Hardie account thus can hardly rely on the premise that most of the predictive EV inferences we draw are in fact good inferences.

Consider now (ii). To support a prediction regarding the effect of P on SAT scores in d_1, Mary will need to figure out whether the required causal principle and interactive factors are present there. And looking at the effect of P in d_0 can be useful for that purpose, even if Argument A is unsound because the relevant causal principles differ or because some of the required support factors differ between d_0 and d_1. For instance, if teacher quality is a support factor for P in d_0, then it would be wise for Mary to check whether it might also

play this role in d_1. In other words, what happened with P in d_0 does not become entirely irrelevant to what will happen with P if implemented in d_1 simply because the predictive EV inference from one to the other would not be a good inference. This means, however, that we should probably stop thinking of the problem of using knowledge garnered from experimental or observational studies to inform policy decisions as being identical with the problem of drawing predictive EV inferences.

7.5 Conclusion

I have argued above that, as far as predictive EV inferences are concerned, the problem of external validity has been solved. The accounts independently developed by Cartwright and Hardie on the one hand and Pearl and Bareinboim on the other converge on the correct answer to the question of the circumstances under which one can infer from the truth of a causal claim in some population to its truth in another population.

If this claim is true, then two consequences immediately follow. First, Steel's demand that any acceptable account of EV inferences break the Extrapolator's Circle is misplaced. Though the demand might be legitimate for methods for drawing EV inferences, it is not for analyses of the circumstances under which such inferences are good. Second, not all EV inferences are inductive, contrary to what Guala and others have claimed. According to the Cartwright-Hardie account, predictive EV inferences are good just in case they are sound deductive arguments of the kind of Argument A.

7.6 Acknowledgments

This chapter has been published in *Philosophy of Science* and is reprinted here with permission from the publisher:

- Marcellesi, A. 2015. "External Validity: Is There Still a Problem?" *Philosophy of*

Science 82(5): 1308-1317.

Chapter 8

Modeling Mitigation and Adaptation Policies to Predict their Effectiveness: The Limits of Randomized Controlled Trials (with Nancy Cartwright)

8.1 Climate Policies: Mitigation and Adaptation

The negative effects of anthropogenic global warming[1] on natural and social systems promise to be diverse and important: melting of glaciers and of the polar ice caps (IPCC, 2007b, 356-360) contributing to a rise of sea-levels (op. cit., 418); increase in the frequency and intensity of extreme weather events like droughts, heat waves, or floods (IPCC, 2012); decrease in crop productivity resulting in increased risk of hunger (IPCC, 2007a, 298); increased risk of extinction for a great number of plant and animal species (op. cit., 792); etc. Most of these negative effects are expected to occur regardless of the way emissions of

[1] We use the expressions 'anthropogenic global warming' and 'climate change' interchangeably in this paper.

greenhouse gases (GHGs) evolve in the future, and some of them are already being observed.

It is not, however, too late for policy makers to act. First, though many of the effects of global warming will inevitably occur, their intensity depends on how large the rise in average temperature turns out to be. Reducing emissions of GHGs, the cause of anthropogenic global warming, can thus help moderate the intensity of these effects. Second, because most of the effects of global warming will inevitably occur, policies for adapting to these effects and limiting their harmful consequences are necessary.[2]

This paper is about some of the serious problems we can expect to face in modeling the effects of climate change policies—in evaluating the effectiveness of policies that have been implemented and in predicting the results of polices that are proposed. The difficulties we will discuss are shared with other kinds of social and economic policies, but they can be particularly problematic for climate change policies, as we will show below. Policies for addressing climate change are commonly divided into two categories, mitigation and adaptation, corresponding to the two levels at which policy makers can address climate change.[3] The Intergovernmental Panel on Climate Change (IPCC) defines a mitigation policy as "A human intervention to reduce the sources or enhance the sinks of greenhouse gases" (IPCC, 2007b, 949) and an adaptation policy as an "Adjustment in natural or human systems in response to actual or expected climatic stimuli or their effects, which moderates harm or exploits beneficial opportunities." (IPCC, 2007a, 869) One can put the distinction between mitigation and adaptation in causal terms by saying that while mitigation policies are designed to reduce the causes of global warming, adaptation policies are designed to moderate its harmful effects on natural and human (or social) systems.

[2]Global warming is expected to have limited positive effects, in the short run and in some regions, for instance in the domain of timber productivity (IPCC, 2007a, 289). It is also the task of policy makers to design policies for taking advantages of these positive effects.

[3]This distinction is reflected in the Fourth IPCC Assessment Report. This report treats of mitigation and adaptation in two distinct parts, though it contains a chapter on the relations between them (IPCC, 2007a, chapter 18).

8.2 Evidence-Based Climate Policies

Agencies which fund mitigation and adaptation policies typically want 'their money's worth'; they want to fund policies 'that work', that is policies that produce the effects they are designed to produce where and when they are implemented.[4] Claims that a given policy 'works', moreover, should be based on evidence. This idea, which is at the root of the widespread evidence-based policy movement, seems natural enough: A policy should be funded, and implemented, only if there is reasonable evidence that it will produce the desired effect in the specific location and at the specific time at which it is implemented.

In order to produce such evidence, organizations implementing policies are invited to conduct 'impact evaluations'. Impact evaluations (IEs) are studies measuring the effects of policy interventions. They are, by definition, retrospective: A policy must have been implemented for its effects to be measured. These IEs have two main functions: First, when an IE establishes that the policy had the effect it was designed to have, it thereby provides a post hoc justification for the decision to fund and implement the policy. Second, the results of IEs are supposed to inform subsequent policy decisions by providing evidence supporting predictions about the effectiveness of policies.

Both functions are important, and this is why many of the agencies that fund policies devote part of their resources to IEs. An example in the domain of climate policies is the Global Environment Facility (GEF). The GEF, an intergovernmental agency which funds many mitigation and adaptation policies, has its own evaluation office, which produces guidelines for conducting IEs.[5]

[4] They also want policies that have large benefit/cost ratios. We leave aside issues related to cost-benefit analysis itself in what follows, and focus on the preliminary step to any such analysis: the evaluation of the likelihood that a policy will yield the intended benefit.

[5] See http://www.thegef.org/gef/eo_office. Other funding agencies such the World Bank (http://ieg.worldbankgroup.org/), the International Monetary Fund (http://www.ieo-imf.org), or the US Food and Drug Administration (http://www.fao.org/evaluation/) also have their own evaluation offices. There are also organizations, such as the International Initiative for Impact Evaluation (3ie, http://www.3ieimpact.org/), whose sole role is to fund and carry out IEs. The multiplication of evaluation offices results in the multiplication of guidelines and methodologies for conducting IEs.

As we mentioned above, the aim of IEs is to measure the effects of policy interventions. This is essentially an issue of causal inference. Teams of researchers that carry out IEs are, in the words of statistician Paul Holland, in the business of "measuring the effects of causes." (Holland, 1986, 945) The extensive literature on causal inference in statistics and related disciplines (e.g. econometrics or epidemiology) provides policy makers with many different methods, experimental and observational, for conducting IEs.

Indeed, the counterfactual approach to causal inference Rubin 1974; Holland 1986 which is prominent in statistics has had a palpable influence on the field of evaluation. According to the World Bank's guide to impact evaluation, for instance,

> To be able to estimate the causal effect or impact of a program on outcomes, any method chosen must estimate the so-called counterfactual, that is, what the outcome would have been for program participants if they had not participated in the program. (Bank, 2011, 8, emphasis added)[6]

As this quotation hints, the idea at the root of the counterfactual approach is that the size of the contribution of a putative cause C to an effect E among program participants is identical to the difference between the value of E for those participants in a situation in which C is present and the value which E would take in a situation in which C is absent, all else being equal. If this difference is equal to zero, then C is not a cause of E in that population; if it is greater than zero, then C is a positive cause of E, and if it is smaller than zero, then C is a negative cause of E. According to the counterfactual approach to causal inference, answering the question 'What is the effect of C on E in a given population?' thus requires answering the following counterfactual queries 'What value would E take for individuals in

[6]It is widely assumed, and not just by the World Bank, that answering a causal question about the effect of a policy just is to answer some counterfactual question about what would have happened in the absence of the policy. Thus Duflo and Kremer, both members of the influential Jameel Poverty Action Lab at MIT, claim that, "Any impact evaluation attempts to answer an essentially counterfactual question: how would individuals who participated in the program have fared in the absence of the program?" (Duflo and Kremer, 2003, 3) And Prowse and Snilstveit, in a review of IEs of climate policies, claim that, "IE is structured to answer the [counterfactual] question: how would participants' welfare have altered if the intervention had not taken place?" (Prowse and Snilstveit, 2010, 233)

that population exposed to C were C absent, all else being equal?' and 'What value would E take for individuals not exposed to C were C present, all else being equal?'

This commitment to a counterfactual approach goes together with a strong preference for experimental methods, and for randomized controlled trials (RCTs) in particular, over observational methods. According to their advocates,[7] RCTs yield the most trustworthy or, as development economists Esther Duflo and Michael Kremer put it (Duflo and Kremer, 2003), "credible" estimates of the mean effect of C on E in a given population. RCTs are, to use a common expression, the 'gold standard' of causal inference.[8]

8.3 What are RCTs, and Why Are They Considered the 'Gold Standard'?

RCTs are experiments in which individuals in a sample drawn from the population of interest are randomly assigned either to be exposed or not exposed to the cause C, where an individual can be anything from a single student to a single village to a hospital to a single country or region. Individuals who are exposed to C form the 'treatment' group while individuals who are not exposed form the 'control' group.[9] Random assignment does, in ideal circumstances and along with a sufficiently large sample, make it probable that the treatment and control groups are homogeneous with respect to causes of E besides C. And the homogeneity of the two groups with respect to causes of E other than C enables one to answer the counterfactual question 'What would be the mean value of E for individuals (in the study population) exposed to C were C absent, all else being equal?' by citing the mean value taken by E for individuals not actually exposed to C.[10] In other words, ideally

[7] Who are sometimes called 'randomistas' as in, e.g., (Ravallion, 2009).

[8] See, e.g., (Rubin, 2008).

[9] The terminology comes from clinical trials.

[10] It also enables one to answer the question 'What would be the mean value of E for individuals (in the study population) not exposed to C were C present, all else being equal?' by citing the mean value taken by E for individuals actually exposed to C. Note that we are here talking about mean values of E over the treatment

conducted RCTs make it likely, by their very design,[11] that all else is indeed equal between the treatment and control groups, and thus that the actual mean value of E for the control group can be identified with the mean value which E would take for the treatment group were individuals in this group not exposed to C (and vice-versa for the control group). This is in turn enables one to estimate the mean of the difference between the effect an individual would have were they subject to C versus were they not—often called the causal or treatment effect of C on E—in the sample, or study population, accurately.[12]

Here is a different way to put it. Assume that the effect of interest E is represented by a continuous variable Y_i and that the putative cause C is represented by a binary variable X_i taking value 1 when individual i is exposed to the cause and 0 when it is not. Assume also that the relationship between X_i and Y_i in the study population is governed by the following linear causal principle:

$$Y_i = a + b_i X_i + W_i \qquad \text{(CP)}$$

Here W_i is a continuous variable which represents factors that are relevant to the value of Y_i besides X_i. And coefficient b_i represents the effect of X_i on Y_i for i. Since b_i represents the individual-level effect of X_i on Y_i, the population-level mean effect of X_i on Y_i is by definition equal to E[b_i], where E[.] is the expectation operator.[13]

Randomly assigning individuals to the treatment and control groups in principle guarantees the probabilistic independence of X_i from both b_i and W_i, and this in turn enables one to accurately estimate E[b_i] from the difference between the expected value of the effect in the treatment group and its expected value in the control group.[14] This difference is equal

and control groups respectively. RCTs enable one to estimate the mean causal effect of C on E in a given population, not the individual causal effect of C on E for any specific individual in this population.

[11] RCTs are, in the words of (Cartwright and Hardie, 2012, Section I.B.5.3), 'self-validating', i.e. their very design guarantees, in ideal circumstances, the satisfaction of the assumptions that must be satisfied in order for the causal conclusions they yield to be true.

[12] For more on RCTs and on the way they establish their conclusions see (Cartwright and Hardie, 2012, Section I.B.5) and (Cartwright, 2010).

[13] We treat 'mean', 'expectation' and 'expected value' as synonyms here.

[14] The probabilistic independence of X_i from bi guarantees that the size of the effect of C on E for i is causally

to:

$$\mathrm{E}[Y_i|X_i=1]-\mathrm{E}[Y_i|X_i=0]=(a+\mathrm{E}[b_i|X_i=1]+\mathrm{E}[W_i|X_i=1]) \\ -(a+\mathrm{E}[b_i|X_i=0]+\mathrm{E}[W_i|X_i=0])$$

In the ideal case in which assignment of individuals to either treatment or control genuinely is independent of b_i and W_i, this difference is the mean treatment effect—often referred to as just the 'treatment effect'—and can be estimated from the observed outcome frequencies. It is equal to:

$$\mathrm{E}[Y_i|X_i=1]-\mathrm{E}[Y_i|X_i=0]=\mathrm{E}[b_i]$$

So the mean treatment effect is non-zero just in case $\mathrm{E}[b_i]$ is non-zero, which can happen only if b_i is non-zero for some i in the population, which means that for that individual X_i does contribute to the value of Y_i: X_i causes Y_i in that i.

Experimental and observational studies in which assignment to the treatment and control groups is non-random are widely considered less desirable than RCTs because their designs, unlike that of RCTs, do not in principle make the causal homogeneity of the two groups (regarding causes of E other than C) probable, even in large samples, or, alternatively, their designs do not guarantee the probabilistic independence of X_i from b_i and W_i. This is why RCTs are considered the 'gold standard' by a large number of social and policy scientists.

If RCTs are the 'gold standard' for measuring the effects of causes, and if the aim of IEs is to measure the effects of policy interventions, then it seems legitimate to conclude that IEs should be designed as RCTs whenever possible. Indeed, this is the view advocated by a variety of policy scientists, for instance members of the Jameel Poverty Action Lab (J-PAL)

---

unrelated to whether i is assigned to the treatment or the control group. And the probabilistic independence of X_i from W_i guarantees that whether i is assigned to the treatment or control group is causally unrelated to the causes of E that do not appear in (CP).

such as Esther Duflo. J-PAL funds and carries out IEs that use RCTs, at the exclusion of any other evaluation methodology.[15] The view that RCTs provide the best evidence regarding the effects of policies is also embraced by the Grading of Recommendations Assessment, Development and Evaluation (GRADE) working group, a group of health scientists that produces standards for rating the quality of evidence. According to GRADE's evidence-ranking scheme, adopted by many agencies worldwide including the World Health Organization, results from RCTs are rated as having 'high quality' while results from observational studies receive a 'low quality' rating (Balshem et al., 2011, 404, table3). The views of these organizations about RCTs are echoed in hundreds of other agencies dedicated to vetting policy evaluations around the Anglophone world in areas from education to crime to aging to climate change.

So are RCTs a "silver bullet" for policy evaluation, to use an expression from (Jones, 2009)? How relevant to policy making is the evidence they generate? Should the evidence base for mitigation and adaptation policies be improved by conducting RCT-based IEs? We will argue below that RCTs have important limitations and that the emphasis put on them contributes to obscuring questions that must be answered for the effectiveness of policy interventions to be reliably predicted. In Sections 8.4 and 8.5 we will show, first in theory and then in practice—using a particular family of mitigation policies as a concrete example, that even if we agree that an RCT is necessary, results from RCTs provide only a small part of the evidence needed to support effectiveness predictions. Then, in Section 8.6, we will show that RCTs are ill-suited to evaluate the effects of most adaptation policies. Our main aim is to underline some particular methodological problems that face the use of RCTs to evaluate mitigation and adaptation policies. We use particular policy examples to illustrate these problems. But we do not aim to offer an exhaustive treatment of these particular policies nor of the full range of challenges that arise in evaluating the effectiveness

[15]Though this does not mean that J-PAL members only work on RCTs, it does mean that all the IEs sponsored and conducted by J-PAL take the form of RCTs.

of mitigation and adaptation policies in general.

8.4 The Limited Relevance of RCTs to Effectiveness Predictions

8.4.1 Internal and External Validity

It is common, in the social and policy sciences, to distinguish between the internal and external validity of studies seeking to measure the effects of causes. According to the standard view, a study is internally valid when it produces results that are trustworthy, and externally valid when its results hold in contexts other than that of the study itself.[16] Because RCTs in principle are supposed to yield the most trustworthy estimates of treatment effects, they are also considered to have the highest degree of internal validity.[17]

It is possible for a study to have a high degree of internal validity while having a very low degree of external validity. A particular RCT, for instance, might yield conclusions that are highly trustworthy but which only hold of the study population involved in the RCT and not of any other population. Results from a study are useful for the purpose of predicting the effectiveness of policy interventions only if they are both internally and externally valid. If IEs are to be useful to policy makers, then, they must produce results that have a high degree of external validity, in addition to being internally valid.

What does it take for a study result to be externally valid? It is often said that, for a study result to hold in contexts other than that of the study itself, the circumstances considered must be 'similar' to that of the study.[18] But what makes a set of circumstances

[16]There is a lot to be said about the standard view and why the labels 'internal validity' and 'external validity' are both vague and misleading. Given limitations of space, however, these issues cannot be discussed here. For more, see (Cartwright and Hardie, 2012, Section I.B.6.3).

[17]The hedge 'in principle' is important. Poorly executed RCTs will not produce trustworthy estimates of treatments effects.

[18]See (Cartwright and Hardie, 2012, Section I.B.6.3) for a concrete example of an appeal to similarity. See

'similar' to some other set of circumstances? We briefly describe a framework, fully developed in (Cartwright and Hardie, 2012), that enables one to address questions of external validity in a rigorous and fruitful manner.

8.4.2 Causal Roles, Causal Principles and Support Factors

Causes do not produce their effects willy-nilly, at least not where it is possible to predict these effects. Rather, the effect of C on E in a given population is governed by *causal principles* that hold in that population. These causal principles can, without real loss of generality, be represented in the form of (CP) above, where C is represented by X_i and E is represented by Y_i.[19] C *plays a causal role* in (CP) just in case it genuinely appears in the equation, i.e. just in case there are values of b_i such that $b_i(X_i = 1) \neq 0$ for some i in the given population. But C does not work alone to produce a contribution to E: It works together with what we call *support factors*. These support factors are represented by b_i in (CP).[20]

The idea that causes work together with support factors derives from the view that causes are INUS conditions in the sense of (Mackie, 1965). To say that C is an INUS condition for E is to say that it is an Insufficient but Necessary part of an Unnecessary but Sufficient condition for the production of a contribution to E.[21] Mackie's classic example is that of a fire caused by a short circuit. The short circuit is not individually sufficient to produce a contribution to the fire, other factors, which we call 'support factors', are required:

---

also http://blogs.worldbank.org/impactevaluations/impactevaluations/why-similarity-wrong-concept-external-validity.

[19] All the conclusions we draw below apply mutatis mutandis when the relevant causal principles take more complex forms than that of (CP) (e.g. non-linear forms).

[20] You may be used to thinking of b_i as the size of the effect of X_i on Y_i. Indeed, this is the way we described it above when introducing (CP). But because, as we explain below, causes are INUS conditions, the two descriptions are equivalent: The effect of C on E just is what happens to E when C is present along with all of its required support factors.

[21] Each term in an equation like (CP) represents a contribution to the effect. Mackie's original theory does not mention 'contributions' because he only consider binary 'yes-no' variables. Our presentation is more general in that it encompasses both cases in which the cause and effect variables are binary, and more common cases in which they are not.

The presence of flammable material, the presence of oxygen, the absence of sprinklers, etc. These support factors, together with the short circuit, are jointly sufficient to produce a contribution to the fire. But they are not jointly necessary: There are other ways to contribute to a fire, i.e. there are other sets of factors—e.g. sets that have lit cigarettes instead of short circuits—that are also jointly sufficient to produce a contribution to the fire.[22]

Policies are causes, and as such are INUS conditions. They generally cannot produce a contribution to the effect they are designed to address by themselves: They need support factors. And the distribution of these support factors will differ from situation to situation. We can even expect considerable variation in which factors *are* support factors, that is which factors are needed to obtain a given effect often varies with context. Consider again Mackie's example as an illustration of this point: The short circuit may not require the absence of sprinklers in houses that are not connected to the water supply system in order to produce a contribution to the fire, though it may require the presence of a particularly large amount of flammable material in houses whose walls have been painted using fire resistant paint in order to produce the same contribution to the fire. There is no 'one size fits all' set of a support factors that, together with the cause of interest, will produce the same contribution to the effect in every context. What matters is the presence of the 'right mix' of support factors, i.e. the presence of the right support factors in the right proportions, and what the 'right mix' consists in often differs from context to context.

The framework briefly sketched above enables one to frame questions about external validity in more precise terms than does the claim that external validity is a matter of how 'similar' sets of circumstances are. To ask whether a trustworthy result from a particular study regarding the mean effect of C on E will hold in a population other than the study population is to ask:

[22] As the 'short circuit' example makes evident, the distinction between policies and support factors is a pragmatic one. Both a policy and its support factors are causes, and so both are INUS conditions. Some factor is usually singled out as the policy because it is practical, ethically acceptable, or cost-efficient to manipulate it. Note also that we claim that all causes are INUS conditions, but not that all INUS conditions are causes.

- Does C play the same causal role in the target population as in the study population?
- Are the support factors required for C to produce a contribution to E present in the right proportions in the target population?

When both questions have positive answers, C will make a positive contribution in the target population if it does so in the study population. If either has a negative answer it is still possible that C will make a positive contribution but the RCT result is irrelevant to predicting whether it will or not—it provides no warrant for such a prediction.

8.4.3 Which Questions do RCTs Answer?

An ideal RCT for the effect of C on E will give you an accurate estimate of E[b_i], the mean value of b_i over individuals in the study population, or treatment effect. If this estimate is larger than 0, then you know that C makes a positive contribution to E for at least some individuals in the study population. And if this estimate is smaller than 0, then you know that C makes a negative contribution to E for at least some individuals in the study population.[23]

An ideal RCT may thus get you started on your external validity inference by providing you with some trustworthy information about the causal role C plays with respect to E in at least one population, the study population. But it gets you nowhere at all towards learning what you need to know about support factors: An ideal RCT will not tell you what the support factors are (i.e. what b_i represents) nor about individual values of b_i, i.e. about the effect of C on E for particular individuals, nor for what proportion of the study population C plays a positive, or negative, role.[24]

[23] If this estimate is equal to 0, or very close to 0, then you cannot directly draw any conclusion about the causal role played by C in the study population because you do not know whether C is ineffective or, alternatively, its positive and its negative effects balance out. We leave this case aside here.

[24] See (Heckman, 1991) for a further critique of the limitations of RCTs when it comes to estimating parameters that are of interest for policy making.

How much further can an ideal RCT can take you on the way to a reliable external validity inference? The short answer is: Not much further. The framework introduced above makes it clear why. First, an ideal RCT will not tell you what the causal principle governing the relationship between C and E in the study population looks like.[25] Second, an ideal RCT will not tell you what the support factors required for C to produce a contribution to E in the study population are, nor how they are distributed. Third, an ideal RCT will not tell you whether C plays the same causal role in the principles governing the production of E in the target population as in the study population. Fourth, an ideal RCT will not give you information about the support factors required for C to produce a contribution to E in the target population, nor about whether the support factors needed in the target population are the same as in the study population (which, very often, is not the case). And you need these pieces of information to produce a reliable prediction about the effectiveness of a policy.

Advocates of RCTs often reply that what is needed to overcome these limitations is more RCTs, but RCTs carried out in different locations.[26] The reasoning underlying this rejoinder seems to be the following: If RCTs conducted in locations A, B, and C all yield conclusive results regarding the effects of a policy, then you have strong evidence that this policy will produce the same effects when you implement it in a fourth location, call it D. This reasoning, however, is problematic insofar as it assumes without justification that the policy can play the same causal role in D as it does in A, B, or C. Since the RCTs in A, B, and C cannot individually tell you what causal principle is at work in each of these locations, their conjunction cannot, a fortiori, tell you what causal principle is at work in D. And if you don't know what causal principle is at work in D, then you also don't know whether the policy can play there the causal role you want it to play.[27]

[25] Apart from giving you a trustworthy estimate of the value of $E[b_i]$.

[26] Banerjee and Duflo, for instance, make the following claim: "A single experiment does not provide a final answer on whether a program would universally 'work'. But we can conduct a series of experiments, differing in [...] the kind of location in which they are conducted..." (Banerjee and Duflo, 2011, 14) They add that, "This allows us to [...] verify the robustness of our conclusions (Does what works in Kenya also work in Madagascar?)..." (ibid.)

[27] You may think this is an uncharitable reconstruction of the argument advanced by advocates of RCTs.

Inferring from results in three—or even a dozen or two dozen—different locations, no matter how different they are, to the next one is a notoriously bad method of inference. It is induction by simple enumeration. Swan 1 is white, swan 2 is white, swan 3 is white... So the next swan will be white. Of course science does make credible inductions all the time. But their credibility depends on having good reason to think that the individuals considered are the same in the relevant way, that is in the underlying respects responsible for the predicted feature. In the case of causal inference from RCT populations that means that they are the same with respect to the causal role C plays and with respect to having the right mix of the right support factors.

Policy scientists writing about mitigation and adaptation policies often lament the current state of the evidence base and, naturally, call for its "strengthening" via rigorous IEs (Prowse and Snilstveit, 2010, 228). So should agencies which fund and implement mitigation and adaptation policies carry out RCTs? Should the GEF, as a report of its Scientific and Technical Advisory Panel urges (STAP, 2010), start designing its policies *as experiments*, and preferably RCTs, in order to improve the evidence base for climate change policies? The discussion above should make it clear that we think that RCTs are of limited relevance when it comes to producing evidence that's relevant for predicting the effectiveness of policies. We illustrate this point in the next section by examining a particular family of mitigation policies.

But the claims they sometimes make, e.g. Banerjee and Duflo's claim, quoted in note 26, regarding the need for several RCTs in order to establish that a policy works "universally", seem to invite reconstructions that are far less charitable. One could thus see advocates of RCTs as advancing an argument of the form 'If RCTs produce conclusive results in A, B, and C, then the policy works "universally", and it will therefore work in D'. This construal seems less charitable in that it attributes to advocate of RCTs a claim (the conditional in the previous sentence) that's highly likely to be false.

8.5 Predicting the Effectiveness of Mitigation Policies

8.5.1 Mitigation Via Payments for Environmental Services

Payment for Environmental Services (PES) programs are policies that seek to conserve the environment by paying landowners to change the way they use their land. Environmental, or ecosystem, services (ESs) are loosely defined as "the benefits people obtain from ecosystems." (MEA, 2005, 26) PES policies involve a buyer, the user of the ES or a third-party acting on her behalf, and a seller, the provider of the ES.[28]

Thus a person who owns a forest and uses it for a timber activity may provide ESs by stopping this activity and by replanting trees that were cut down. In this case, the ESs provided consist in the protection of currently existing carbon stocks, via avoided deforestation, and the improvement of carbon sequestration, via the planting of new trees. Both of these ESs are directly relevant to climate change mitigation, though not all PES programs target ESs that are relevant to climate change mitigation. Many PES programs are designed with the conservation of biodiversity as their main aim.[29]

In order to stop her timber activity, the landowner described above must have an incentive to do so. Why stop her timber activity if this means a loss of earnings, and why replant trees if this means a cost without a benefit? This is where PES programs come in: They are supposed to create the incentives necessary for landowners to change the way they use their land and provide an ES. As Engel et al. put it: "The goal of PES programs is to make privatively unprofitable but socially-desirable practices become profitable to individual land users, thus leading them to adopt them." (Engel et al., 2008, 670)[30]

[28] In the case of mitigation-relevant PES program, the buyer of the ES often is an intergovernmental agency, e.g. the GEF, acting as a third-party on behalf of users of the ES. When the GEF is the buyer of the ES, the users it represents are the citizens of states that are members of the UN.

[29] Of course, many PES programs that target biodiversity also results in the protection of carbon stocks and, conversely, many PES programs that target climate change mitigation also result in the conservation of biodiversity.

[30] The theory behind PES programs comes from the work of Ronald Coase on social cost (Coase, 1960). But see (Muradian et al., 2010) for an alternative theoretical framework within which to understand PES programs.

Governmental and intergovernmental agencies see PES programs targeting deforestation as offering a major opportunity for mitigating climate change. A significant portion of the total emissions of GHGs, and CO_2 in particular, comes from deforestation.[31] If PES programs can create incentives to reduce deforestation, especially in developing tropical countries in which deforestation is a major concern, then they can contribute to a reduction in emissions of GHGs, and thus to a moderation of global warming and of its negative effects.[32]

PES programs are modeled after existing conditional cash transfer programs in domains such as development, for instance the Mexican *Oportunidades* program.[33] There are numerous IEs, including ones that take the form of RCTs, measuring the effects of conditional cash transfer programs that target poverty-reduction and education. This is particularly true for the *Oportunidades* program, first implemented in 1997 (See, e.g., Parker and Teruel 2005). This is not the case for PES programs and, in particular, for those PES programs that are relevant to climate change mitigation. There are few IEs measuring the effects of PES programs on, e.g., deforestation. And there are no completed IEs of PES programs that takes the form of an RCT.

The current state of the evidence base for PES programs is deplored by Pattanayak et al., who "see an urgent need for quantitative causal analyses of PES effectiveness. (Pattanayak et al., 2010, 267) "Such analyses", they add, "would deliver the hard numbers needed to give policy makers greater confidence in scaling up PES." (ibid.) In this spirit, the report to the GEF mentioned above (STAP, 2010) urges the intergovernmental organization

[31] 20% according to (IPCC, 2007b), 12% according to (van der Werf et al., 2009).

[32] The UN, for instance, is developing a program called 'REDD+' that relies on PES-type programs in order to reduce deforestation. Note that 'REDD' is an acronym for 'Reduction of (carbon) Emissions from Deforestation and forest Degradation'.

[33] In the *Oportunidades* (originally PROGRESA) program, parents receive conditional payments for activities that improve human capital, e.g., enrolling their children to school. The idea is to reduce poverty both in the short-term, via the cash payments, and the in the long-run, by improving human capital. The payments in this program, as well as in PES programs, are conditional in that they are made only if the service (e.g. an ES) is actually provided: They are not one-time payments that are made upfront.

to design its policies—including PES programs—as experiments as much as is possible, and this in order to facilitate the evaluation of their effects.

8.5.2 What Will RCTs Add to the Evidence Base for PES Programs?

Responding to the call for an improvement of the evidence base for the effectiveness of PES programs in securing environmental services, MIT's J-PAL, in collaboration with the International Initiative for Impact Evaluation (3ie) and Innovations for Poverty Action (IPA), is currently (as of August 2013) carrying out an RCT aimed at measuring the effectiveness of a PES program in reducing deforestation and biodiversity loss in the Hoima and Kibaale districts of Western Uganda.[34] Deforestation rates are particularly high in these two districts, where landowners "often cut trees to clear land for growing cash crops such as tobacco and rice or to sell the trees as timber or for charcoal production." (Jayachandran, 2013a)

The design of J-PAL's RCT is as follows (Jayachandran, 2013b, 311). First, 1,245 private forest owners—spread over 136 villages—were identified. They form the RCT's study population. A survey was then conducted to record several of their characteristics: number of hectares of land owned, past tree-cutting behavior, attitude toward the environment, access to credit, etc. 65 out of the 136 villages—representing 610 landowners—were then randomly assigned to the treatment group, the remaining villages being assigned to the control group. Landowners residing in villages in the treatment group were called into meetings by a local non-governmental organization (NGO), the Chimpanzee Sanctuary & Wildlife Conservation Trust (CSWCT), to receive information about the program as well as contract forms. The 'treatment' that is randomly assigned in this RCT can thus be described as 'Being offered the opportunity to sign a PES contract with CSWCT'. One of the aims pursued by J-PAL's scientists here is to estimate the effect of this treatment on deforestation and biodiversity loss.

[34]The project was supposed to last for four years, from April 2010 through April 2014. As of January 2016, it remains 'under implementation' and results are not available yet.

Landowners who chose to participate in the program (or take up the 'treatment') then signed contracts with the local NGO. As Jayachandran (ibid.) reports,

> The contract specifies that the forest owner will conserve his entire existing forest, plus has the option to dedicate additional land to reforestation. Under the program, individuals may not cut down medium-sized trees and may only cut selected mature trees, determined by the number of mature trees per species in a given forest patch. Participants are allowed to cut small trees for home use and to gather firewood from fallen trees.

Compliance with the contract is monitored via spot checks by CSWCT staff. Landowners who comply receive $33/hectare of forest preserved annually, an amount that was selected because it is assumed to be greater than what landowners would earn from cutting down and selling trees (other than those specified by the PES contract) for timber or charcoal, or from clearing land to grow cash crops (e.g. tobacco). As we indicated above, the assumption guiding the design of this and other PES programs is that agents will modify their behavior—here, will stop cutting down trees—if they are given the right monetary incentives to do so.

This RCT, as the official project description states, is justified by the fact that "although many PES schemes have been undertaken globally, there has not been concrete proof, emanating from scientific empirical data collected from real life PES schemes, that they are effective." (GEF, 2010, 6) Note, furthermore, that this study is funded by the GEF, whose administration thus seems to be sensitive to the call for RCT-based IEs of PES programs that can deliver "hard numbers" and give "concrete proof" based on "scientific empirical data" of the effectiveness of "real life" PES programs.

As the project description indicates, one of the aims of the study is to generate, develop and disseminate a "replicable PES model based on lessons learned and best practices." (GEF, 2010, 3) The aim of this RCT thus is not simply to demonstrate the effectiveness of the specific PES programs implemented in the Hoima and Kibaale districts in producing ESs. The explicit aim is to show that PES programs aimed at reducing deforestation and

biodiversity loss are effective in general, and to develop a PES model that can be scaled up and applied in locations besides select districts in Western Uganda.

Is the RCT currently carried out by J-PAL likely to achieve the result sought? Is it likely to provide strong evidence that PES programs work in general? How much evidence can it provide for this conclusion? If you are a policy maker contemplating the implementation of a PES program, is the RCT likely to provide reasonably strong evidence that such a program will work in the location you are targeting? We do not believe so, for reasons that were advanced in their theoretical form in Section 8.4.3. The J-PAL RCT, if it is carried out according to the script, will deliver an accurate estimate of the mean effect of the PES program on deforestation and biodiversity loss in the study population.

But it will not reveal the causal principle governing the relationship between the PES program and the reduction of deforestation and biodiversity loss in the study population.[35] It also won't tell you what support factors are needed for the PES program to play a positive causal role in the study population, nor how these factors are distributed in this population. The J-PAL RCT will not, a fortiori, tell you where the causal principle at work in the study population also holds in the population you are targeting. And it won't tell you what the support factors required for the PES program to play a positive causal role in the target population are, nor how they will be distributed.

One needs these essential additional pieces of information, regarding causal principles and support factors, in order to predict at all reliably whether the PES program will play the same causal role when it is implemented in other locations, e.g. when it is scaled up to other districts in Western Uganda, or when it is implemented in Eastern Uganda, or when it is implemented in other countries in sub-Saharan Africa, etc. One cannot arrive at a "replicable PES model", i.e. at a PES model that will work in many locations, without a detailed understanding of how the PES program works in the original study population.

[35] And it won't tell you whether the same causal principle is at work in those parts of the study populations composed of landowners from the Hoima district and those parts composed of landowners the Kibaale districts.

Nor is it clear that there is a reliable "replicable PES model" that works 'in general' to be found. It is not obvious that one can formulate substantial and useful generalizations about PES programs across settings (cultural, political, economic, religious, etc.) and, especially, across types of ESs (Can one generalize results obtained in a context in which the ES is avoided deforestation to a context in which the ES is the preservation of water resources?). The framework introduced above is designed to help you think about how a policy works when it does, and about what it would take for it to work in a different location.

We are obviously not claiming that nothing will have been learned during the four years of the J-PAL project described above, besides an estimate of some treatment effect. The policy scientists carrying out J-PAL's RCTs are neither blind nor stupid. They will gain a wealth of new knowledge regarding the local institutional and social context, the way landowners respond to the PES program, differences between villages that are relevant to the effect of the program, etc. Note, however, that this context-specific knowledge (1) may well have been acquired even if enrollment in the PES program had not been randomly offered to landowners, (2) is just as important as is knowledge of the treatment effects to predicting the effectiveness of subsequent PES programs, and (3) is likely to be overshadowed by the "hard numbers", i.e. the estimates of treatment effects. The framework introduced above, and fully developed in (Cartwright and Hardie, 2012), shows why this context-specific knowledge is essential to predicting the effectiveness of policies. And it also gives you the tools to articulate this knowledge in ways that make it relevant to effectiveness predictions.

The bottom line, here, is that if you are a policy maker contemplating the implementation of a PES program for reducing deforestation and biodiversity loss in a particular location, the results from J-PAL's RCT will offer you some guidance, but not much. You need knowledge about the causal principles at work and the support factors required for the PES program to produce a positive contribution in the location you are targeting. Let us further illustrate the importance of support factors by looking at five hypothesized support

factors needed by PES programs in some locations.

8.5.3 Some of the Support Factors (Sometimes) Needed by PES Programs

We briefly list below five of the factors identified in the literature as playing a role in determining the effectiveness of PES programs in reducing deforestation and biodiversity loss.[36] As we noted above in section 8.4.2, a policy might require different support factors in different contexts in order to produce the intended contribution to the effect of interest. These five factors, therefore, may be support factors for PES programs in some contexts, but not in others. The second factor—the low cost of enforcing PES programs—for instance, may not be a required support factor in contexts in which the sellers of the ES tend to abide by contracts for cultural or religious reasons.

Our framework makes it plain why these factors matter and why having evidence about their presence and distribution is crucial. If we make the unrealistic assumption that these factors are support factors always required by PES programs then, for your effectiveness prediction regarding a PES program to be properly supported by evidence, you must have evidence that these factors are present, and distributed in just the right way, in the location in which the program is to be implemented.[37] Below we list the five factors we have seen cited in the literatures about PES programs and some of the questions they immediately give rise to. But behind these there are bigger questions that need answering: 'Are these necessary in all cases?', 'What else is necessary in any particular case?', 'Will the necessary factors be in place, or can they be put in place, in the new place?', and very importantly, 'What kinds of study can help us find out the answers to these bigger questions?'

[36] See e.g. (Pattanayak et al., 2010), (Pirard et al., 2010), (Alix-Garcia et al., 2009), (GEF, 2010, 35) or (Jayachandran, 2013b).

[37] And if the assumption that these factors are always required is dropped, then you also need evidence that these factors are indeed support factors needed for the PES program to produce the intended contribution to the effect in the location you are targeting.

1. *Strong property rights*: A PES program, it is argued, can only be effective if there exists property rights and the means to enforce them in the location in which the program is to be implemented. There is no landowner for the ES buyer to sign a contract with if there is no landowner to start with. But how strong do these property rights need to be, and do they need to be guaranteed by a government? Where are property rights strong enough, and where are they too weak for PES programs to be effective?

2. *Low cost of monitoring and enforcing PES contracts*: If the economic and political cost of monitoring and enforcing PES contracts is high then there is an incentive for the buyer not to do so, and thus for the seller to breach the contract. These costs must be low for PES programs to be effective. But how low must they be? And how does one assess these costs?

3. *Sustainable and flexible funding source*: PES programs can only be effective, it is argued, if they are funded on the long-term and if the funding source is flexible enough to allow for re-negotiation of PES contracts. If the price of timber rises, then the payment for forest conservation provided to a forest owner must rise for the incentives to stay the same, and for the forest owner to keep providing an ES. Can NGOs provide sustainable and flexible funding? What about governmental agencies in countries that are politically unstable?

4. *Absence of leakage*: If a forest owner agrees to stop her timber activity on a parcel she owns and for which the PES contract was signed, but then goes on to use the extra earnings from the contract to buy a similarly-sized parcel nearby and resume her timber activity on that parcel, then the PES program is not effective in reducing deforestation and biodiversity loss. Opportunities for 'leakage' must be limited for the PES program to play the expected causal role. How does one assess opportunities

for leakage?

5. *Access to credit*: If a forest owner cannot easily borrow money to cover emergency expenses (e.g. medical bills), then she might cut down and sell trees instead, even if she signed a PES contract covering those trees. An easy access to credit might thus lower the chances that forest resources will be used as a 'safety net' and thus have a bearing on the effectiveness of the PES program. But how exactly does one measure 'access to credit', and how easy must access to credit be in order for the resources covered by the PES contract to stop being a 'safety net'?

We emphasize that these are just five among the numerous factors that may be support factors required for a PES program to produce a contribution to the reduction of deforestation. The point we want to illustrate here is that J-PAL's RCT will not tell you whether these are support factors required in the location you are targeting, nor whether they are actually present there, nor how they are distributed. Unfortunately, you need this information in order to accurately predict whether a PES program will play the causal role you want it to play in the location in which you are contemplating its implementation.

8.6 Evaluating the Effects of Adaptation Policies: The Limits of RCTs.

Remember that adaptation policies seek to modify natural or human systems in order to reduce their vulnerability to weather-related events due to climate change. The term 'vulnerability' has a precise meaning in this context. According to the IPCC's definition, the vulnerability of a system (usually some geographical unit, e.g. a city) to climate change is the "degree to which [it] is susceptible to, and unable to cope with, adverse effects of climate change, including climate variability and extremes." (IPCC, 2007a, 883) More precisely, the vulnerability of a system is "a function of the character, magnitude, and rate of climate

change and variation to which [it] is exposed, its sensitivity, and its adaptive capacity." (ibid.) An adaptation policy is designed to reduce the vulnerability of a system by reducing its sensitivity—i.e. the extent to which it is harmed by climate change—or by enhancing its adaptive capacity—i.e. its ability to adjust to moderate the harmful effects of climate change. A distinction is often drawn between environmental vulnerability—as measured for instance by the country-level Environmental Vulnerability Index (EVI)—and social vulnerability—as measured for instance by one of the Social Vulnerability Indices (SoVi).[38]

There are various obstacles to the use of RCT-based IEs to evaluate the effects of adaptation policies. First, adaptation policies take a wide variety of forms, many of which simply do not lend themselves to randomization. Consider for instance the 'Adaptation to Climate Change through Effective Water Governance' policy under implementation in Ecuador that aims to improve the country's adaptive capacity by mainstreaming "climate change risks into water management practices…" (GEF, 2007, 2) This policy will change water management practices in Ecuador, e.g. by incorporating climate risks in the country's 'National Water Plan'. How is one to evaluate the extent to which such a policy will improve Ecuador's adaptive capacity and thus reduce its vulnerability, both environmental and social, to climate change? RCTs are no help here, given that the policy is implemented at the level of an entire country. One cannot, for a variety of reasons (political, practical, etc.), randomly assign countries to particular policy regimes.

The same point applies to the many adaptation policies that aim to improve some country's adaptive capacity, and thus reduce its vulnerability, by modifying its institutions. Here is another example. The government of Bhutan is, with the help of the United Nations Development Programme (UNDP), implementing the 'Reducing Climate Change-Induced

[38] See http://www.vulnerabilityindex.net/ for the EVI and http://webra.cas.sc.edu/hvri/ for the US county-level SoVI. Note two difficulties with using these indices to evaluate the effects of adaptation policies. First, they are measures of vulnerability to environmental hazards in general, whether or not they are due to climate change. Second, there is no wide consensus as to how to measure overall vulnerability (at various geographical scales), and neither is there a consensus regarding how to measure an important component of vulnerability, namely adaptive capacity.

Risks and Vulnerabilities from Glacial Lake Outburst Floods [GLOFs]' policy which, among other things, aims to integrate the risk of GLOFs due to climate change occurring in the Punakha-Wangdi and Chamkhar valleys in Bhutan's national disaster management framework.[39] Such policies, because they target country-level institutions, cannot in practice be evaluated using RCT-based IEs. The problem here is that a vast number of adaptation policies fall into this category. Note also that such policies, by their very nature, are tailored to the institutions of a particular country and so may not be implementable in any other country. A policy that improves Bhutan's adaptive capacity, for instance, may not be applicable, and a fortiori may not have the same beneficial effects, in a country which faces similar risks but has a different institutional structure (e.g. Canada, which, unlike Bhutan, is a federal state).

Second, for many adaptation policies, RCT-based IEs are superfluous. Consider for instance the Kiribati Adaptation Program (Phase II) implemented between 2006 and 2010 that included the construction of a 500 meters long seawall to protect the country's main road, a coastal road around Christmas Island. One does not need an RCT in order to determine whether this seawall is helping protect the road and reduce beach erosion (inside this wall). The physical configuration of seawalls guarantees that they will reduce the sensitivity of the systems inside them to the consequences of climate change (e.g. to rising sea levels, erosion, and extreme weather events). One might argue that an RCT would enable one to determine by how much the Kiribati seawall reduces the sensitivity of the systems it helps protects, i.e. would enable one to estimate the size of the effect of this seawall on sensitivity. In this case, as with most adaptation policies, however, the need for an immediate reduction in sensitivity trumps the need for precise estimates of treatment effects.

One could have conducted an RCT in which the coastline along the Christmas Island

[39] See http://www.adaptationlearning.net/bhutan-reducing-climate-change-induced-risks-and-vulnerabilities-glacial-lake-outburst-floods-punakh.

road is divided into *n* sections, half of them randomly assigned to the 'seawall' group and half of them to the 'no seawall' group, and compared the condition of the road and the extent of beach erosion between sections in the 'seawall' group and those in the 'no seawall' after a year, for instance. This would have provided one with estimates of the effect of seawalls on road condition and beach erosion on Kiribati's Christmas Island (assuming both road condition and beach erosion can be reliably measured). Conducting such an RCT would make little sense for Kiribati's policy makers, however. Roads are useful only if they enable you to get somewhere, and they can only do so if they are uninterrupted and in good condition rather than irreversibly damaged at random intervals. The aim of this hypothetical example is not to caricature the position of those who, like members of the GEF's Scientific and Technical Advisory Panel (STAP, 2010), call for more RCT-based IEs of adaptation and mitigation policies. It is simply to illustrate that such calls sometimes conflict with the goals the policies that are to be evaluated are supposed to achieve. What matters in the end is that these policies produce the beneficial effects they were designed to produce, not that we have highly trustworthy point estimates of the size of these effects.

This is not to say that there are no adaptation policies the effects of which can be evaluated using RCT-based IEs. Policies which offer farmers rainfall index insurance, i.e. policies that insure farmers against both deficits and excesses in rainfall, can be considered adaptation policies, and their effects on the vulnerability of particular study populations to climate change can in principle be evaluated using RCTs, even though no such RCT has been conducted to date.[40] This is true in general of adaptation policies that do not seek to reduce a country's vulnerability by modifying its institutions (e.g. by incorporating climate risks into its planning tools) or its infrastructures (e.g. by building seawalls) but rather target units (e.g. individual farmers or villages) that can more easily be randomly assigned to some

[40]RCTs conducted about weather insurance usually attempt to estimate the effects of such insurance on investment decisions (see e.g. Giné and Yang 2009) or to understand the causes of weather insurance take-up (see e.g. Cole et al. 2013). See (de Nicola, 2011) for a non-randomized evaluation of the effects of rainfall index insurance on the welfare of farmers and so on their adaptive capacity.

treatment group. The mistake here would be to think that such policies should occupy a privileged position in the portfolio of policies available to policy makers preoccupied with adapting to climate change simply because they can be evaluated using RCT-based IEs. As we showed in Section 8.5 for PES policies aiming at mitigation, the fact that a policy lends itself to randomization does not imply that it can more easily be generalized beyond the study population. And it also does not imply that this policy is more effective than other policies that cannot be similarly evaluated. A policy that offered Ugandan farmers the possibility of using drought-resistant seeds might lend itself to an RCT-based IE more easily than does Uganda's national irrigation masterplan, but this obviously does not mean that the former is more effective than the latter at reducing the sensitivity of Ugandan farmers to droughts due to climate change.

We showed in Section 8.5 that results from RCT-based IEs of mitigation policies such as PES programs provide only a small part of the total evidence needed to support effectiveness predictions. The situation is more challenging even in the case of adaptation policies, since many of these cannot be evaluated using RCTs in the first place. The lesson of this section thus is that, both for evaluating past adaptation policies and for supporting predictions regarding the effectiveness of future adaptation policies, we need more than RCTs. Nor is it especially the issue of random assignment that raises difficulties. We face here rather problems that are endemic with comparative group studies: They are often not possible and they tell us only a little of what we need to know to make use of their own results.

8.7 Conclusion

Should J-PAL scientists pack their bags and cancel the RCT they are currently carrying out in Western Uganda? No. Are RCTs a bad tool for causal inference? No. Are estimates of treatment effects irrelevant for policy making in the domain of climate change

policies? No.

We want to emphasize that our criticisms are not directed at RCTs per se. Criticizing RCTs in principle makes little more sense than criticizing hammers in principle. Both RCTs and hammers are well-designed tools. One can criticize their instances: There are bad hammers and poorly conducted RCTs. And one can criticize the use to which they are put. It is the use to which RCTs are frequently put that we target and criticize.

Calling for more and more RCTs in order to strengthen the evidence base for mitigation policies such as PES programs is a bit like calling for the use of more and more hammers in order to carve a statue out of a block of marble. What one needs is not more and more hammers, but hammers and chisels, i.e. tools of a different kind. In the policy case, what one needs is not more estimates of treatment effects produced by more RCTs. If one starts with an RCT, what one needs is evidence of a different kind, evidence that is relevant to external validity inferences, and so to prediction about the effectiveness of particular policies implemented in particular contexts. The framework sketched above in Section 8.4.2 tells you what kind of evidence is needed, namely evidence about causal principles and support factors.

What we advocate corresponds, to some extent, to what (Pattanayak et al., 2010, 6) call "economic archeology", i.e. the qualitative evaluation of existing policies in order to reveal the contextual factors that are relevant to their effectiveness. What we argue is that calls for an improvement of the evidence base for PES programs, and mitigation and adaptation policies in general, should emphasize the need for more "economic archeology" just as much, or even more, than they emphasize the need for estimates of treatment effects generated by RCTs. This is particularly true for adaptation policies since, as we showed in Section 8.6, these often cannot be evaluated using RCTs. The "hard numbers" produced by RCTs—when and where they are available—are of little use for policy without knowledge of the networks of factors that give rise to these numbers, and without models of these

networks (See Cartwright forthcoming). The framework sketched here, and fully developed in (Cartwright and Hardie, 2012), provides one with the means to do ”economic archeology” where RCTs are involved in a rigorous and fruitful manner.

But it is important to stress that we do not need to start with RCTs in order to pursue economic archeology. The issue of course is how to do economic archeology in anything like a rigorous and reliable way. This involves understanding how best we can provide evidence about causal relations in the single case. So, besides a call for more and more RCTs, surely there should be an equally urgent call for more systematic study of what counts as evidence for causality in the single case.

8.8 Acknowledgments

This chapter was coauthored with Nancy Cartwright and is included in this dissertation with her permission:

- Marcellesi, A. and Cartwright, N. 2013. “Modeling Climate Mitigation and Adaptation Policies to Predict their Effectiveness: The Limits of Randomized Controlled Trials.”

Bibliography

Alexander, Joshua. 2012. *Experimental Philosophy: An Introduction*. Polity Press.

Alix-Garcia, Jennifer, de Janvry, Alain, Sadoulet, Elisabeth, and Torres, Juan Manuel. 2009. "Lessons Learned from Mexico's Payment for Environmental Service Program." In Leslie Lipper, Takumi Sakuyama, Randy Stringer, and David Zilberman (eds.), *Payment for Environmental Services in Agricultural Landscapes*, volume 31 of *Natural Resource Management and Policy*, 163–188. New York: Springer.

Angrist, Joshua and Pischke, Jörn-Steffen. 2009. *Mostly Harmless Econometrics*. Princeton University Press.

Bacon, Francis. 1878[1620]. *Novum Organum*. Oxford: Clarendon Press.

Balshem, Howard, Helfand, Mark, Schünemann, Holger J., Oxman, Andrew D., Kunz, Regina, Brozek, Jan, Vist, Gunn E., Falck-Ytter, Yngve, Meerpohl, Joerg, Norris, Susan, and Guyatt, Gordon H. 2011. "{GRADE} guidelines: 3. Rating the quality of evidence." *Journal of Clinical Epidemiology* 64:401 – 406. ISSN 0895-4356. doi: http://dx.doi.org/10.1016/j.jclinepi.2010.07.015.

Banerjee, Abhijit and Duflo, Esther. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty*. New York: PublicAffairs.

Bank, World. 2011. *Impact Evaluation in Practice*. Washington, D.C.: World Bank.

Bareinboim, Elias and Pearl, Judea. 2013. "A General Algorithm for Deciding Transportability of Experimental Results." *Journal of Causal Inference* 1:107–134.

Bateman, Ian, Munro, Alistair, Rhodes, Bruce, Starmer, Chris, and Sugden, Robert. 1997. "A Test of the Theory of Reference-Dependent Preferences." *Quarterly Journal of Economics* 112:479–505.

Baumgartner, Michael. 2009. "Interdefining Causation and Intervention." *Dialectica* 63:175–194.

—. 2012. "The Logical Form of Interventionism." *Philosophia* 40:751–761.

—. 2013. "A Regularity Theoretic Approach to Actual Causation." *Erkenntnis* 78:85–109. doi:10.1007/s10670-013-9438-3.

Bigaj, Tomasz. 2012. "Causation Without Influence." *Erkenntnis* 76:1–22.

Blank, Rebecca, Dabady, Marilyn, and Citro, Constance (eds.). 2004. *Measuring Racial Discrimination*. Panel on Methods for Assessing Discrimination. Washington, D.C.: The National Academies Press.

Briggs, Rachael. 2012. "Interventionist Counterfactuals." *Philosophical Studies* 160:139–166. doi:10.1007/s11098-012-9908-5.

Campbell, Donald. 1957. "Factors Relevant to the Validity of Experiments in Social Settings." *Psychological Bulletin* 54:297–312.

Campbell, John. 2007. "An Interventionist Approach to Causation in Psychology." In Alison Gopnik and Laura Schulz (eds.), *Causal Learning: Psychology, Philosophy and Computation*, 58–66. Oxford University Press.

Carnap, Rudolf. 1946. "Modalities and Quantification." *The Journal of Symbolic Logic* 11:33–64.

Cartwright, Nancy. 1979. "Causal Laws and Effective Strategies." *Noûs* 13:419–437.

—. 2007. *Hunting Causes and Using Them*. Cambridge: Cambridge University Press.

—. 2010. "What are randomised controlled trials good for?" *Philosophical Studies* 147:59–70.

—. 2014. "Single Case Causes: What is Evidence and Why."

—. forthcoming. "Will Your Policy Work? Experiments vs. Models." In Bas van Fraassen and Isabelle Peschard (eds.), *The Experimental Side of Modeling*. Chicago: University of Chicago Press.

Cartwright, Nancy and Hardie, Jeremy. 2012. *Evidence-Based Policy: A Practical Guide to Doing it Better*. New York: Oxford University Press.

Coase, Ronald. 1960. "The Problem of Social Cost." *Journal of Law and Economics* 3:1–44.

Cole, Shawn, Giné, Xavier, Tobacman, Jeremy, Topalova, Petia, Townsend, Robert, and Vickery, James. 2013. "Barriers to Household Risk Management: Evidence from India." *American Economic Journal: Applied Economics* 5:104–135.

Collingwood, Robin. 1940. *An Essay on Metaphysics*. Oxford: Clarendon Press.

Cook, Thomas and Campbell, Donald. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Houghton Mifflin.

Craver, Carl. 2007. *Explaining the Brain*. Oxford University Press.

de Nicola, Francesca. 2011. "The Impact of Weather Insurance on Consumption, Investment, and Welfare." Technical Report 548, Society for Economic Dynamics.

Divers, John. 2002. *Possible Worlds*. London: Routledge.

Dowe, Phil. 2000. *Physical Causation*. Cambridge: Cambridge University Press.

Duflo, Esther and Kremer, Michael. 2003. "Use of randomization in the evaluation of development effectiveness." .

Dupré, John. 1984. "Probabilistic Causality Emancipated." *Midwest Studies In Philosophy* 9:169–175.

Eberhardt, Frederick and Scheines, Richard. 2007. "Interventions and Causal Inference." *Philosophy of Science* 74:981–995.

Engel, Stefanie, Pagiola, Stefano, and Wunder, Sven. 2008. "Designing Payments for Environmental Services in Theory and Practice: An Overview of the Issues." *Ecological Economics* 65:663–674.

Franklin-Hall, Laura. forthcoming. "High-Level Explanation and the Interventionist's 'Variables Problem'." *British Journal for the Philosophy of Science* .

Frege, Gottlob. 1879. *Begriffsschrift, ein der arithmetischen nachgebildete Formelsprache des reinen Denkens*. Halle/Saale: Verlag L. Nebert. Translated as *Concept Script, a formal language of pure thought modelled upon that of arithmetic*, by S. Bauer-Mengelberg in J. vanHeijenoort (ed.), *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*, Cambridge, MA: Harvard University Press, 1967, pp. 1–82.

Frisch, Ragnar and Waugh, Frederick. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica* 1:387–401.

Gasking, Douglas. 1955. "Causation and Recipes." *Mind* 64:479–487.

GEF. 2007. "Adaptation to Climate Change through Effective Water Governance in Ecuador." Project executive summary, Global Environment Facility, Washington, D.C.

—. 2010. "Developing an Experimental Methodology for Testing the Effectiveness of Payments for Ecosystem Services to Enhance Conservation in Productive Landscapes in Uganda." Technical report, Global Environment Facility, Washington, D.C.

Giné, Xavier and Yang, Dean. 2009. "Insurance, Credit, And Technology Adoption : Field Experimental Evidence From Malawi." *Journal of Development Economics* 89:1–11.

Glymour, Clark. 1986. "Comment: Statistics and Metaphysics." *Journal of the American Statistical Association* 81:964–966.

—. 2004. "Critical notice of Woodward, *Making Things Happen*." *British Journal for the Philosophy of Science* 55:779–790.

Glymour, Clark, Danks, David, Glymour, Bruce, Eberhardt, Frederick, Ramsey, Joseph, Scheines, Richard, Spirtes, Peter, Teng, Choh Man, and Zhang, Jiji. 2010. "Actual Causation: A Stone Soup Essay." *Synthese* 175:169–192.

Greiner, James and Rubin, Donald. 2011. "Causal effects of perceived immutable characteristics." *The Review of Economics and Statistics* 93:775–785.

Guala, Francesco. 2005. *The Methodology of Experimental Economics*. Cambridge University Press.

Hall, Ned. 2004. "Two Concepts of Causation." In John Collins, Ned Hall, and Laurie Paul (eds.), *Causation and Counterfactuals*. Cambridge: MIT Press.

—. 2007. "Structural Equations and Causation." *Philosophical Studies* 132:109–136.

—. Ms. "Structural Equations and Causation (extended version)." .

Halpern, Joseph and Hitchcock, Christopher. 2014. "Graded Causation and Defaults." *British Journal for the Philosophy of Science* 66:413–457.

Hardimon, Michael. 2012. "The Idea of a Scientific Concept of Race." *Journal of Philosophical Research* 37:249–282.

Harker, David. 2012. "Inference to the Best Explanation and the Importance of Peculiarly Explanatory Virtues."

Harman, Gilbert. 1965. "The Inference to the Best Explanation." *Philosophical Review* 74:88–95.

Heckman, James. 1991. "Randomization and Social Policy Evaluation." *NBER Working Paper Series* 107.

Hempel, Carl and Oppenheim, Paul. 1948. "Studies in the Logic of Explanation." *Philosophy of Science* 15:133–175.

Hitchcock, Christopher. 1995. "Salmon on Explanatory Relevance." *Philosophy of* 62:304–320.

—. 2001. "The Intransitivity of Causation Revealed in Equations and Graphs." *Journal of Philosophy* 98:273–299.

—. 2007. "Prevention, Preemption, and the Principle of Sufficient Reason." *Philosophical Review* 116:495–532.

—. 2009. "Structural Equations and Causation: Six Counterexamples." *Philosophical Studies* 144:391–401.

Hitchcock, Christopher and Woodward, James. 2003. "Explanatory Generalizations, Part 2: Plumbing Explanatory Depth." *Noûs* 37:181–199.

Holland, Paul. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81:945–960.

—. 2003. “Causation and Race.” Technical Report RR-03-03, Educational Testing Services.

Hoover, Kevin. 2004. “Lost Causes.” *Journal of the History of Economic Thought* 26:149–164.

Illari, Phyllis and Russo, Federica. 2014. *Causality. Philosophical Theory meets Scientific Practice*. Oxford: Oxford University Press.

IPCC. 2007a. *Climate Change 2007: Impacts, Adaptation and Vulnerability*. New York: Intergovernmental Panel on Climate Change.

—. 2007b. *Climate Change 2007: The Physical Science Basis*. New York: Intergovernmental Panel on Climate Change.

—. 2012. *Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation*. New York: Intergovernmental Panel on Climate Change.

Jayachandran, Seema. 2013a. “Evaluating a Payments for Ecosystem Services program in Uganda.” Blog post on www.climate-eval.org.

—. 2013b. “Liquidity Constraints and Deforestation: The Limitations of Payments for Environmental Services.” *American Economic Review* 103:309–313.

Jones, Harry. 2009. “The ‘gold standard’ is not a silver bullet for evaluation.” *Overseas Development Institute Opinion* 127.

Kahneman, Daniel, Knetsch, Jack, and Thaler, Richard. 1990. “Experimental Tests of the Endowment Effect and the Coase Theorem.” *Journal of Political Economy* 98:1325–1348.

Kuorikoski, Jaakko. 2014. “How to Be a Humean Interventionist.” *Philosophy and Phenomenological Research* 89:333–351.

Kvart, Igal. 2001. “Lewis’s ‘Causation as Influence’.” *Australasian Journal of Philosophy* 79:409–421.

Kyburg, Henry. 1965. “Discussion: Salmon’s Paper.” *Philosophy of Science* 32:147–151.

LaFollette, Hugh and Shanks, Niall. 1995. “Two Models of Models in Biomedical Research.” *Philosophical Quarterly* 45:141–160.

Lewis, David. 1973a. “Causation.” *Journal of Philosophy* 70:556–567.

—. 1973b. *Counterfactuals*. Oxford: Blackwell.

—. 1979. “Counterfactual Dependence and Time’s Arrow.” *Noûs* 13:455–76.

—. 1983. "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61:343–77.

—. 1986a. "Causal Explanation." In *Philosophical Papers, Volume 2*. Oxford: Oxford University Press.

—. 1986b. *On the Plurality of Worlds*. Oxford: Basil Blackwell.

—. 1986c. "Postscripts to 'Causation'." In *Philosophical Papers*, volume II, 159–213. Oxford: Oxford University Press.

—. 2000. "Causation as Influence." *Journal of Philosophy* 97:182–197.

List, Christian and Menzies, Peter. 2009. "Non-reductive physicalism and the limits of the exclusion principle." *Journal of Philosophy* 106.

Mackie, J.L. 1965. "Causes and Conditions." *American Philosophical Quarterly* 2:245–264.

Mackonis, Adolfas. 2013. "Inference to the Best Explanation, Coherence and Other Explanatory Virtues." *Synthese* 190:975–995.

MEA. 2005. *Ecosystems and Human Well-Being: Synthesis (A Report of the Millennium Ecosystem Assessment)*. Washington, D.C.: Island Press.

Menzies, Peter. 2004. "Difference-making in context." In John Collins, Ned Hall, and Laurie Paul (eds.), *Causation and Counterfactuals*, 139–180. Cambridge, MA: MIT Press.

—. 2012. "The causal structure of mechanisms." *Studies in History and Philosophy of Biological and Biomedical Sciences* 43:796–805.

—. 2014. "Counterfactual Theories of Causation." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*. Spring 2014 edition.

Menzies, Peter and Price, Huw. 1993. "Causation as a Secondary Quality." *British Journal for the Philosophy of Science* 44:187–203.

Milkowski, Marcin. 2013. *Explaining the Computational Mind*. Cambridge, MA: MIT Press.

Mill, John Stuart. 1843. *A System of Logic*.

Muradian, Roldan, Corbera, Esteve, Pascual, Unai, Kosoy, Nicolas, and May, Peter. 2010. "Reconciling theory and practice: An alternative conceptual framework for understanding payments for environmental services practice: An alternative conceptual framework for understanding payments for environmental services." *Ecological Economics* 69:1202–1208.

Nolan, Daniel. 2005. *David Lewis*. Chesham: Acumen Publishing.

Parker, Susan and Teruel, Graciela. 2005. "Randomization and Social Program Evaluation: The Case of Progresa." *The ANNALS of the American Academy of Political and Social Science* 599:199–219.

Pattanayak, Subhrendu, Wunder, Sven, and Ferraro, Paul. 2010. "Show Me the Money: Do Payments Supply Environmental Services in Developing Countries?" *Review of Environmental Economics and Policy* 4:254–274.

Paul, Laurie and Hall, Ned. 2013. *Causation: A User's Guide*. New York: Oxford University Press.

Pearl, Judea. 2000. *Causality: Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.

Pirard, Romain, Billé, Raphaël, and Sembrés, Thomas. 2010. "Questioning the theory of Payments for Ecosystem Services (PES) in light of emerging experience and plausible developments." *Analyses (Biodiversity)* 4:5–22.

Prowse, Martin and Snilstveit, Birte. 2010. "Impact evaluation and interventions to address climate change: a scoping study." *Journal of Development Effectiveness* 2:228–262.

Ragin, Charles. 1987. *The Comparative Method*. University of California Press.

Ravallion, Martin. 2009. "Should the Randomistas Rule?" *The Economists' Voice* 6:1–5.

Reiss, Julian and Cartwright, Nancy. 2005. "Uncertainty in Econometrics: Evaluating Policy Counterfactuals." In Peter Mooslechner, Helene Schuberth, and Martin Schurz (eds.), *Economic Policy Under Uncertainty: The Role Of Truth And Accountability In Policy Advice*. Edward Elgar Publishing.

Reutlinger, Alexander. 2013. *A Theory of Causation in the Social and Biological Sciences*. Palgrave Macmillan.

Rubin, Donald. 1974. "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies." *Journal of Educational Psychology* 66:688–701.

—. 1986. "Comment: Which ifs have causal answers?" *Journal of the American Statistical Association* 81:961–962.

—. 2008. "Comment: The Design and Analysis of Gold Standard Randomized Experiments." *Journal of the American Statistical Association* 103:1350–1353.

Salmon, Wesley. 1971. "Statistical Explanation." In Wesley Salmon (ed.), *Statistical Explanation and Statistical Relevance*, 29–87. Pittsburgh: University of Pittsburgh Press.

Schaffer, Jonathan. 2000. "Trumping Preemption." *Journal of Philosophy* 97:165–181.

—. 2001. "Causation, Influence, and Effluence." *Analysis* 61:11–19.

Spirtes, Peter, Glymour, Clark, and Scheines, Richard. 2001. *Causation, Prediction, and Search*. Cambridge: MIT Press, 2nd edition.

STAP. 2010. "Payments for Environmental Services and the Global Environment Facility: A STAP Advisory Document." Technical report, Global Environment Facility: Scientific and Technical Advisory Panel, Washington, D.C.

Starmer, Chris. 2000. "Developments in Non-Expected Utility Theory: The Hunt for a Descriptive Theory of Choice Under Risk." *Journal of Economic Literature* XXXVIII:332–382.

Steel, Daniel. 2008. *Across The Boundaries*. Oxford University Press.

Stone, Jim. 2009. "Trumping the Causal Influence Account of Causation." *Philosophical Studies* 142:153–160.

Strevens, Michael. 2003. "Against Lewis's New Theory of Causation: A Story With Three Morals." *Pacific Philosophical Quarterly* 84:398–412.

—. 2004. "The Causal and Unification Approaches to Explanation Unified—Causally." *No* 38:154–176.

—. 2007. "Review of Woodward, *Making Things Happen*." *Philosophy and Phenomenological Research* LXXIV:233–249.

—. 2008a. "Comments on Woodward, *Making Things Happen*." *Philosophy and Phenomenological Research* LXXVII:171–192.

—. 2008b. *Depth*. Cambridge, MA: Harvard University Press.

Suppes, Patrick. 1970. *A Probabilistic Theory of Causality*. Amsterdam: North-Holland Publishing Company.

van der Werf, G, Morton, D, DeFries, R, Olivier, J, Kasibhatla, P, Jackson, R, Collatz, G, and Randerson, J. 2009. "CO_2 Emissions from Forest Loss." *Nature Geoscience* 2:737–738.

von Wright, Georg. 1971. *Explanation and Understanding*. Ithaca: Cornell University Press.

Weslake, Brad. 2010. "Explanatory Depth." *Philosophy of Science* 77:273–294.

Woodward, James. 2000. "Explanation and Invariance in the Special Sciences." *British Journal for the Philosophy of Science* 51:197–254.

—. 2003a. "Critical Notice: *Causality* by Judea Pearl." *Economics and Philosophy* 19:321–340.

—. 2003b. *Making Things Happen*. New York: Oxford University Press.

—. 2004. “Counterfactuals and Causal Explanation.” *International Studies in the Philosophy of Science* 18:41–72.

—. 2008. “Response to Strevens.” *Philosophy and Phenomenological Research* LXXVII:193–212.

—. 2010. “Causation in Biology: Stability, Specificity, and the Choice of Levels of Explanation.” *Biology and Philosophy* 25:287–318.

—. 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* 81:691–713.

Woodward, James and Hitchcock, Christopher. 2003. “Explanatory Generalizations, Part 1: A Counterfactual Account.” *Noûs* 37:1–24.

Yablo, Stephen. 1992. “Mental Causation.” *Philosophical Review* 101:245–80.